Electronic Supplement to System One Decision Models: A Survey of the Typed Decision Contract, Its Ecosystem and Its Evidence

AMIR RAFE, PH.D., Texas State University, USA
SUBASISH DAS, PH.D., Texas State University, USA

This electronic supplement accompanies the survey of typed probabilistic decision models. It gives the proofs of the propositions, the search record, the risk-of-bias grade of every study, the card checklist, the census detail, the pattern catalog and the evidence ledger.

Additional Key Words and Phrases: System One models, decision models, evidence ledger, risk of bias, census

S1 Proofs of the propositions

This part of the electronic supplement gives the proofs of the three propositions of Section 5 whose arguments take more than a few lines. The proofs of Propositions 5.1, 5.2 and 5.3 are short and appear in the main text. All probabilities below are taken over the population on which the contract is evaluated, and the same arguments hold in a finite sample when every probability is replaced by the corresponding empirical share.

Proof of Proposition 5.4. The recall identity is the multiplication rule of conditional probability applied to the events G and H given $Y = 1$, and it needs no assumption about how the two stages depend on each other. The cost of one item is the random quantity $C = c_1 + \mathbf{1}_{\text{esc}}\, c_2 + \mathbf{1}_{\text{rev}}\, c_{\text{R}} + L_{\text{res}}$, since the screening stage always runs, and linearity of expectation gives the stated expected cost without any independence among its terms.

Let p be a calibrated probability, so that $\Pr(Y = 1 \mid p) = p$. Flagging an item then has expected cost $(1 - p)\, c_{\text{FP}}$ and passing it has expected cost $p\, c_{\text{FN}}$, so flagging is optimal exactly when $(1 - p)\, c_{\text{FP}} \le p\, c_{\text{FN}}$, which rearranges to $p \ge t^*$ [54].

A review that returns the correct label at cost c_{R} adds a third action whose expected cost is c_{R} for every p. The optimal action minimizes the three expected costs, as in the reject option of Chow [34], so review is optimal exactly when $c_{\text{R}} \le p\, c_{\text{FN}}$ and $c_{\text{R}} \le (1 - p)\, c_{\text{FP}}$. The inequalities are equivalent to $p \ge c_{\text{R}}/c_{\text{FN}}$ and $p \le 1 - c_{\text{R}}/c_{\text{FP}}$, which is the band (7). The band is non-empty exactly when $c_{\text{R}}/c_{\text{FN}} \le 1 - c_{\text{R}}/c_{\text{FP}}$, and multiplying by $c_{\text{FP}} c_{\text{FN}}$ gives $c_{\text{R}}(c_{\text{FP}} + c_{\text{FN}}) \le c_{\text{FP}} c_{\text{FN}}$, with positive width exactly when the inequality is strict.

The lower edge satisfies $c_{\text{R}}/c_{\text{FN}} \le t^*$ exactly when $c_{\text{R}} \le c_{\text{FP}} c_{\text{FN}}/(c_{\text{FP}} + c_{\text{FN}})$, and the upper edge satisfies $1 - c_{\text{R}}/c_{\text{FP}} \ge t^*$ under the same condition, so a non-empty band always contains t^*. Outside the band the rule reduces to the two-action rule, passing items below $c_{\text{R}}/c_{\text{FN}}$ and flagging items above $1 - c_{\text{R}}/c_{\text{FP}}$. Under calibration the expected cost of the rule is $\mathbb{E}[\min(p\, c_{\text{FN}}, (1 - p)\, c_{\text{FP}}, c_{\text{R}})]$ and the expected share of items sent to review is $\Pr(c_{\text{R}}/c_{\text{FN}} \le p \le 1 - c_{\text{R}}/c_{\text{FP}})$, and both depend on the distribution of p alone.

The union bound is Boole's inequality, which holds for events with any dependence. When the failure events are independent with probabilities e_j, the probability of at least one failure equals $1 - \prod_j (1 - e_j)$, which never exceeds $\sum_j e_j$. Either expression is an identity or a bound in terms of the rates $\Pr(D_j)$ under the deployed workflow, so it becomes an estimate only when those rates are measured on the inputs that each component receives inside that workflow. □

Authors' Contact Information: Amir Rafe, Ph.D., Texas State University, San Marcos, USA, amir.rafe@txstate.edu; Subasish Das, Ph.D., Texas State University, San Marcos, USA, subasish@txstate.edu.

Proof of Proposition 5.5. Locality means that the logit of an option is a function $\ell(n, r)$ of its own name and rubric once the state and the question are fixed, and every pair in $\mathcal{N} \times \mathcal{R}$ is realized by some assignment of names to rubrics. The decomposition (8) is the two-way decomposition of the array ℓ, with μ its grand mean, $\alpha(r)$ the mean over names minus μ, $\beta(n)$ the mean over rubrics minus μ, and γ the remainder. If two such decompositions existed, their difference would be a centered decomposition of the zero array, and averaging it over all cells, over names and over rubrics shows in turn that every term of the difference is zero.

An assignment is a bijection σ from rubrics to names, and the returned distribution identifies only log-odds, which remove μ and any constant shared by the options of one query. The log-odds of rubric r against rubric r' equal $\ell(\sigma(r), r) - \ell(\sigma(r'), r')$. If σ gives name n to r and n' to r' and the swap exchanges them, the change in these log-odds equals $2[\beta(n) - \beta(n')] + \gamma(n, r) + \gamma(n, r') - \gamma(n', r) - \gamma(n', r')$. For a binary question $\gamma(n, r) + \gamma(n, r')$ is the sum of γ over all rubrics for the name n, which is zero, and the same holds for n'. For general K, averaging the interaction contrast over all ordered pairs of distinct rubrics gives a multiple of $\sum_r \gamma(n, r) - \sum_r \gamma(n', r)$, which is zero, so the average change identifies $\beta(n) - \beta(n')$.

Let the names be random, exchangeable and independent of the rubrics. The log-odds of r against r' are $\ell(N_r, r) - \ell(N_{r'}, r')$, and because N_r and $N_{r'}$ share one marginal distribution, their expectation is $\mathbb{E}_N[\ell(N, r)] - \mathbb{E}_N[\ell(N, r')]$, a function of the rubrics alone. Exchangeability also gives the swapped pair $(N_{r'}, N_r)$ the same joint distribution as $(N_r, N_{r'})$, so a swap has no effect on the distribution of the output.

For name invariance, write $f(n, r) = \beta(n) + \gamma(n, r)$, so that under σ the logit of rubric r is $\mu + \alpha(r) + f(\sigma(r), r)$. Two logit vectors give the same softmax distribution exactly when they differ by a constant, so invariance holds exactly when, for a fixed reference σ_0 and every σ, the difference $f(\sigma(r), r) - f(\sigma_0(r), r)$ is the same for every rubric r. If $\beta \equiv 0$ and $\gamma \equiv 0$ then $f \equiv 0$ and invariance holds. For the converse with $K > 2$, set $G(n, r) = f(n, r) - f(\sigma_0(r), r)$, which invariance makes constant along every bijection. Any two cells (n, r) and (n', r') with $n \neq n'$ and $r \neq r'$ lie on a common bijection, so G takes one value on them. Two cells in one row or one column share a value with a third cell outside both their rows and both their columns, which exists because $K > 2$, so G is a constant b.

It follows that $f(n, r) = f(\sigma_0(r), r) + b$ depends on r alone. Averaging f over rubrics gives $\beta(n)$, because γ is centered over rubrics, so β is constant and, being centered, equal to zero. Then $\gamma(n, r)$ depends on r alone, and centering over names makes it zero. For a binary question the centering leaves one free interaction value g, with $\gamma = g$ on the two cells of one assignment and $\gamma = -g$ on the two cells of the other. The log-odds under the two assignments are then $\Delta\alpha + \Delta\beta$ and $\Delta\alpha - \Delta\beta$, in which g cancels, so the outputs coincide exactly when $\Delta\beta = 0$, and g is unidentified and changes no output. □

Proof of Proposition 5.6. The expected level is a linear map from the probability simplex on the levels, whose dimension is $L - 1$, to the interval $[1, L]$. For $L > 2$ this map is many-to-one, so there are distinct distributions A and B with $m(A) = m(B)$, and the mean does not determine the cumulative probabilities. Let a model issue A for every item while the outcomes follow B. Then $\mathbb{E}[Y \mid m(\hat{p})] = m(B) = m(A)$, so the expected level is calibrated, while $A \neq B$ implies that the cumulative probabilities of A and B differ at some $k < L$, where ordinal calibration fails. When $L = 2$ the mean determines the distribution and the two notions coincide.

The worked example of Section 5 is the extreme case in which every cumulative probability fails. Forecast A places 0.8 on the middle level and 0.1 on each neighbor, forecast B places 0.5 on each extreme level, and both have expected level 3.0. The cumulative probability of B equals 0.5 at every $k < L$, while the cumulative probabilities of A are 0, 0.1, 0.1 + 0.8 and 1, none of which equals 0.5.

For strict propriety, let the outcome follow a distribution q with cumulative probabilities Q_k. For every threshold k the expected squared difference satisfies $\mathbb{E}[(\hat{F}_k - \mathbf{1}\{Y \le k\})^2] = (\hat{F}_k - Q_k)^2 + Q_k(1 - Q_k)$, which is the Brier score of the threshold event $\{Y \le k\}$ [15]. The expected RPS is the average of these terms over $k < L$, so it is minimized exactly when $\hat{F}_k = Q_k$ for every k, and because the cumulative probabilities determine the distribution, the unique minimizer is $\hat{p} = q$ [55, 64].

A program that acts when the level exceeds k consumes the probability $1 - \hat{F}_k$ of that event, and by Proposition 5.4 its Bayes threshold is guaranteed optimal when $\Pr(Y > k \mid \hat{F}_k) = 1 - \hat{F}_k$, which is ordinal calibration at k. Neither calibration of the expected level, shown above to be insufficient, nor top-label calibration, which concerns the modal level alone, implies this property. The vendor numbers levels from zero, which shifts every expected level by one and leaves every statement above unchanged. □

S2 Search strings, sweep log and screening counts

Table S1 lists every arXiv query of the sweep of Section 2 in run order with its hit count and retries. A query that received no feed is marked as failed, not as zero hits, and is listed again as a rerun. Sixteen identifiers from the seed list and early screening were force included and are listed in the candidate file of the companion repository. Citation chasing from the launch post and two early Jev evaluations found no citing works on OpenAlex or Semantic Scholar by the cutoff.

Screening against the definition of Section 2 included 243 of the 271 candidate arXiv records and excluded 28 that matched a query term but did not concern a typed decision model. Grey literature added 8 evaluations, for 248 studies in the ledger of Section 7. On the cutoff date, a rerun of eight queries hit a rate limit and is logged as unresolved. A second rerun, labeled curl rerun, used a different HTTP client the next day and answered all eight. It returned three new records dated by the cutoff, two of which entered the ledger, and its hits are in the totals above.

The v1.2 sweep reran every query from 14 September 2026 to 28 October 2026, added queries for models released after the first cutoff, paged through every result and returned 249 new records. These were screened on title and abstract, extracted in batches of about ten under the extraction brief of the companion repository and graded on full text under the anchors of Section S3. A script confirmed that each excerpt appears verbatim in the saved full text and contains its value or, for a table, that the caption does and the value appears in the paper. Every computed value quotes the numbers it uses, each number entering the synthesis carries a selection line naming the ledger row, column and reason, and each accuracy contrast carries a coded comparator class.

The ecosystem census searched Hugging Face and GitHub with terms such as system-one, decision model, typed decision, jev, laya, kev, nimble and decider, each query logged in the companion repository. Hugging Face gave 513 unique entries created from 15 September 2026, and direct lookups raised this to 522. Of these, 101 are original implementations with their own decision head, 259 are derivatives such as conversions, quantizations and fine-tunes, 54 are datasets and 108 are exclusions, mostly name collisions or empty cards. GitHub gave 226 repositories, all 226 created on or after the release date of the hosted model. They comprise 54 integrations, 48 local servers, 47 applications, 36 with open-model training code, 25 evaluation packages, 15 directories and 1 mirror. A concurrent analysis reports many more GitHub projects because it collects every repository that calls the hosted model [120], whereas this screen kept only those its search terms returned. The v1.2 census added 2,645 Hugging Face entries and 15,788 GitHub repositories created by 28 October 2026, the GitHub total including older repositories that the first sweep missed by capping results per query.

Table S1. Every arXiv API query of the two literature sweeps, with its hit count and any retries. Queries marked [v1.2] belong to the second sweep, over submissions from 14 September to 28 October 2026.

Query	Hits	Retries	Outcome
`all:Jev`	16	0	OK
`all:"System One model"`	3	0	OK
`all:"System One models"`	3	0	OK
`all:"decision model" AND all:typed`	3	0	OK
`all:"typed decision"`	5	0	OK
`all:Laya AND all:calibrat*`	–	4	No feed after retries, rerun below
`all:Kev AND all:decision`	1	0	OK
`all:Nimble AND all:decision`	–	4	No feed after retries, rerun below
`all:TypeSafe`	–	4	No feed after retries, rerun below
`all:"decision models" AND all:calibration`	2	0	OK
`abs:"typed probabilistic"`	–	4	No feed after retries, rerun below
`all:"decision model" AND all:Jev`	6	0	OK
`all:jev-1.13`	–	4	No feed after retries, rerun below
`all:RLCD`	–	4	No feed after retries, rerun below
`all:"calibrated decisions"`	1	0	OK
`all:Laya AND all:calibrat* [curl rerun]`	0	3	OK; 0 not previously screened
`all:Nimble AND all:decision [curl rerun]`	0	0	OK; 0 not previously screened
`all:TypeSafe [curl rerun]`	2	0	OK; 2 not previously screened
`abs:"typed probabilistic" [curl rerun]`	0	0	OK; 0 not previously screened
`all:"jev-1.13" [curl rerun]`	0	0	OK; 0 not previously screened
`all:RLCD [curl rerun]`	2	0	OK; 2 not previously screened
`all:"System One" AND all:decision [curl rerun]`	6	0	OK; 2 not previously screened
`all:"decision models" AND all:LLM [curl rerun]`	2	0	OK; 1 not previously screened
`all:Jev [v1.2]`	182	0	OK; 164 not previously screened [v1.2]
`all:"System One model" [v1.2]`	51	0	OK; 47 not previously screened [v1.2]
`all:"System One models" [v1.2]`	51	0	OK; 47 not previously screened [v1.2]
`all:"decision model" AND all:typed [v1.2]`	66	0	OK; 61 not previously screened [v1.2]
`all:"typed decision" [v1.2]`	58	0	OK; 53 not previously screened [v1.2]
`all:Laya AND all:calibrat* [v1.2]`	0	0	OK; 0 not previously screened [v1.2]
`all:Kev AND all:decision [v1.2]`	10	0	OK; 9 not previously screened [v1.2]
`all:Nimble AND all:decision [v1.2]`	4	0	OK; 4 not previously screened [v1.2]
`all:TypeSafe [v1.2]`	23	0	OK; 21 not previously screened [v1.2]
`all:"decision models" AND all:calibration [v1.2]`	61	0	OK; 58 not previously screened [v1.2]
`abs:"typed probabilistic" [v1.2]`	1	0	OK; 1 not previously screened [v1.2]
`all:"decision model" AND all:Jev [v1.2]`	89	0	OK; 82 not previously screened [v1.2]
`all:"jev-1.13" [v1.2]`	6	0	OK; 6 not previously screened [v1.2]
`all:RLCD [v1.2]`	9	0	OK; 7 not previously screened [v1.2]
`all:"calibrated decisions" [v1.2]`	18	0	OK; 15 not previously screened [v1.2]
`all:"System One" AND all:decision [v1.2]`	58	0	OK; 52 not previously screened [v1.2]
`all:"decision models" AND all:LLM [v1.2]`	58	0	OK; 54 not previously screened [v1.2]
`all:"jev-1.14" [v1.2]`	0	0	OK; 0 not previously screened [v1.2]
`all:"Decision-1" AND all:Microsoft [v1.2]`	0	0	OK; 0 not previously screened [v1.2]
`all:Clef AND all:decision [v1.2]`	5	0	OK; 5 not previously screened [v1.2]
`all:"pplx-decider" [v1.2]`	1	0	OK; 1 not previously screened [v1.2]
`all:"Trio-Spark" [v1.2]`	0	0	OK; 0 not previously screened [v1.2]
`all:"System-1" AND all:"decision model" [v1.2]`	12	0	OK; 12 not previously screened [v1.2]
`all:"decision models" AND all:typed [v1.2]`	66	0	OK; 61 not previously screened [v1.2]

S3 Risk-of-bias ratings by study

The risk-of-bias instrument for the ledger of Section 7 adapts the domain logic of PROBAST [250] and QUADAS-2 [249] to typed probabilistic decision models and adds the concerns raised by QUADAS-AI [206] and STARD-AI [205]. Its five domains are reference-label independence, sampling and reporting, version pinning and run dates, tuning-budget parity and test-set exposure, each rated low, unclear or high with a one-line justification from the full text or the extraction note.

Both sweeps were rated a second time, independently of the first. For the first sweep, a computer scientist outside the author team screened all 22 arXiv candidates and agreed on 21, with a kappa of 0.78. The same rater checked 40 randomly sampled ledger values against their sources, 38 of which matched, and rated the domains of a random sample of studies, agreeing on 30 of 40 ratings with a weighted kappa of 0.57. An earlier round without written anchors reached a weighted kappa of 0.46, after which we wrote the anchors in the protocol of the companion repository from the reasons recorded with the first ratings. The other 19 first-sweep studies were later rated by 5 further raters blind to the first rating, who also checked one extracted value per study against its source. They agreed on 60 of 95 domain ratings, with a weighted kappa of 0.44, and on 18 of 19 values. For the second sweep, 10 second raters blind to the first rating together screened all 249 candidates and rated the five domains for all 215 graded studies. They also checked 294 extracted values, at least one per study, and coded the decision head of 36 census originals. They agreed with the first rating on 235 screening decisions, with a kappa of 0.70, on 793 of 1066 domain ratings, with a weighted kappa of 0.63, on 293 values and on 33 head codes, with a kappa of 0.87. The weighted kappa was highest for sampling and reporting, at 0.74, and test-set exposure, at 0.71. It was lowest for reference-label independence, at 0.46, and tuning-budget parity, at 0.51, the two domains that ask how a study built its labels and chose its settings.

We adjudicated every disagreement of both rounds against the full text, which in the first round changed six ratings and one extracted value. In the second round we first wrote rules for the cases the anchors left open, now stated in the protocol, and then decided the 273 disputed domain ratings under them. Of these, 149 kept the first rating and 124 changed, moving the overall grade of 30 studies. Adjudication also added 6 studies that the first screening had excluded, which were then extracted and graded like the rest. It changed 2 head codes, and applying the same rule to every census original changed 1 more. The later first-sweep round was decided under the same rules plus one for model names without a dated snapshot, and 22 of its 35 disputed ratings changed. The 6 studies added at adjudication were then rated by 3 further raters, who agreed on 23 of 30 domain ratings, with a weighted kappa of 0.62, and on 6 of 6 values. Adjudication kept the first rating in all 7 disputed cells. Every graded study, 248 in all, therefore has an independent second rating, and the counts in this section describe the adjudicated ratings.

The overall grade follows a fixed rule applied mechanically. A study is graded low when all five domains are low, or when exactly one is unclear and none is high. It is graded high when reference-label independence or test-set exposure is high, or when any two domains are high, and some concerns otherwise. Under this rule, 3 of the 248 studies are graded low, 103 some concerns and 142 high. Our own studies [179–181] are graded by the same rules as every other study.

The grades are most sensitive to how public benchmarks are rated on test-set exposure. High is reserved for verbatim items with answer keys from openly distributed benchmarks released before 2026, since the training data of the hosted model are undisclosed and the study ran no contamination check. Reading every such rating as unclear grades 3 studies low, 184 some concerns and 61 high. Among those moving from high to some concerns are Li [104] and Yu and Yao [272], whose only high rating is on exposure, and Zhang et al. [276], whose other high rating is on

tuning-budget parity. The choice thus regrades several of the most carefully designed measurement studies but no vendor or practitioner source.

Table S2 gives the domain ratings and overall grade of every graded study, citing grey-literature sources under their main-text keys, and Table S3 gives the justification recorded for each high rating.

Table S2. Risk-of-bias ratings for every ledger study on D1 reference labels, D2 sampling and reporting, D3 version pinning, D4 tuning parity and D5 test-set exposure, each ● high, ◑ unclear or ○ low.

Study	D1	D2	D3	D4	D5	Overall
Ahmed [1]	◑	○	●	●	●	high
Amit [2]	◑	◑	◑	◑	●	high
Antunes et al. [4]	◑	○	●	○	◑	some concerns
Asseri and Asseri [5]	◑	○	○	○	●	high
Azizi et al. [6]	◑	◑	◑	○	◑	some concerns
Azizi et al. [7]	●	◑	●	●	◑	high
Bailo [8]	●	○	●	◑	○	high
Bellver-Soler et al. [11]	◑	○	●	●	●	high
Besanson [12]	●	○	◑	○	○	high
Brom [16]	◑	○	◑	◑	●	high
Bru et al. [17]	◑	○	◑	◑	◑	some concerns
Buecheler et al. [18]	○	○	●	○	○	some concerns
Bukhari et al. [19]	◑	◑	◑	●	◑	some concerns
Cao et al. [21]	◑	◑	◑	◑	●	high
Chen [22]	○	○	◑	◑	◑	some concerns
Chen [24]	○	○	◑	◑	◑	some concerns
Chen and Li [23]	●	○	◑	○	○	high
Chen et al. [27]	◑	●	●	○	●	high
Chen [25]	◑	○	●	◑	◑	some concerns
Chen et al. [26]	◑	◑	●	◑	●	high
Chen et al. [30]	◑	○	◑	◑	◑	some concerns
Chen and Le [29]	◑	○	○	◑	◑	some concerns
Chen et al. [28]	◑	◑	◑	◑	◑	some concerns
Cheng et al. [31]	◑	◑	●	◑	●	high
Cheng et al. [32]	●	◑	●	●	○	high
Chouhan and Perry [33]	◑	○	●	●	◑	high
Chung et al. [35]	◑	◑	◑	◑	●	high
Cinà [36]	○	○	●	●	◑	high
Dai et al. [40]	◑	◑	●	◑	◑	some concerns
Dai et al. [41]	●	◑	●	◑	○	high
Deka [42]	◑	○	◑	◑	◑	some concerns
Demir [43]	◑	○	◑	○	◑	some concerns
Deng et al. [44]	◑	◑	◑	○	○	some concerns
Denney and DiGiuseppe [45]	◑	○	○	◑	●	high
Deußer et al. [46]	◑	○	◑	◑	◑	some concerns
Ding et al. [47]	◑	◑	◑	●	◑	some concerns
Ding [48]	○	◑	◑	◑	●	high
Dinzinger et al. [49]	◑	◑	◑	◑	●	high
Dong et al. [50]	●	◑	◑	●	○	high
Rosario and Atef [189]	○	○	◑	●	●	high

Study	D1	D2	D3	D4	D5	Overall
dos Santos Barbosa [51]	○	●	●	●	◑	high
Du et al. [52]	◑	◑	◑	○	●	high
Duan et al. [53]	◑	◑	●	◑	●	high
Fadli et al. [56]	◑	○	◑	◑	●	high
Feng et al. [57]	◑	○	◑	◑	●	high
Fernandes and Pereira [58]	◑	◑	◑	○	◑	some concerns
Gan et al. [60]	◑	○	◑	◑	●	high
Gao et al. [61]	◑	◑	◑	○	●	high
Gershon et al. [62]	◑	○	○	○	◑	some concerns
Geuter et al. [63]	●	○	◑	●	◑	high
Guo et al. [66]	◑	○	◑	◑	●	high
Guo et al. [65]	◑	○	◑	●	◑	some concerns
Ha et al. [67]	◑	◑	◑	●	●	high
Han et al. [69]	○	◑	◑	◑	◑	some concerns
Han et al. [68]	●	◑	◑	◑	○	high
He et al. [70]	◑	○	◑	○	◑	some concerns
Hu et al. [72]	●	○	◑	◑	◑	high
Hu et al. [73]	◑	○	◑	◑	◑	some concerns
Huang et al. [76]	◑	○	◑	◑	◑	some concerns
Huang et al. [75]	◑	◑	●	◑	●	high
Huang et al. [74]	◑	○	◑	◑	◑	some concerns
Huo et al. [78]	●	◑	●	◑	○	high
Ibrahim and Zaki [80]	○	○	●	●	●	high
Ichikawa [81]	◑	◑	◑	◑	◑	some concerns
Imperatrice et al. [82]	○	○	○	○	◑	low
Iqbal et al. [83]	◑	○	○	◑	◑	some concerns
Jiang et al. [84]	●	●	●	◑	●	high
Joy [85]	◑	○	◑	◑	●	high
Jung [86]	○	○	◑	◑	◑	some concerns
Khatri et al. [87]	◑	○	●	○	●	high
Khezresmaeilzadeh et al. [88]	◑	○	◑	◑	◑	some concerns
Konak [89]	◑	○	◑	◑	◑	some concerns
Kourtis and Xilouris [91]	●	◑	◑	◑	○	high
Kwartler et al. [92]	◑	◑	◑	○	○	some concerns
Lai et al. [93]	◑	○	◑	◑	◑	some concerns
Lai et al. [94]	●	◑	◑	◑	○	high
Lan et al. [96]	◑	◑	◑	◑	◑	some concerns
Lan et al. [95]	◑	◑	●	◑	●	high
Lee et al. [97]	◑	◑	◑	◑	●	high
Lee and Park [98]	○	○	◑	●	◑	some concerns
Li [111]	◑	◑	◑	◑	●	high
Li et al. [106]	○	○	◑	○	◑	some concerns
Li et al. [114]	◑	○	◑	◑	●	high
Li and Xie [108]	◑	◑	◑	●	◑	some concerns
Li et al. [109]	◑	◑	●	◑	○	some concerns
Li et al. [101]	○	◑	◑	●	○	some concerns
Li [105]	◑	●	◑	○	●	high
Li and Wang [107]	◑	◑	◑	◑	●	high
Li et al. [116]	◑	○	◑	●	◑	some concerns

Study	D1	D2	D3	D4	D5	Overall
Li and Chen [110]	◑	◑	◑	●	●	high
Li [104]	○	○	◑	◑	●	high
Li and Wagle [117]	◑	◑	◑	●	◑	some concerns
Li et al. [115]	●	◑	●	◑	◑	high
Li et al. [102]	◑	◑	◑	○	○	some concerns
Li et al. [118]	◑	○	◑	◑	○	some concerns
Li et al. [113]	◑	○	◑	●	●	high
Li [103]	◑	◑	◑	●	●	high
Li et al. [100]	●	○	●	◑	○	high
Li et al. [112]	◑	◑	●	◑	●	high
Li et al. [99]	●	◑	◑	◑	◑	high
Lin [119]	●	○	◑	◑	◑	high
Ling et al. [121]	○	○	◑	◑	◑	some concerns
Liu [127]	●	○	○	◑	◑	high
Liu et al. [124]	○	◑	◑	◑	●	high
Liu et al. [128]	◑	○	●	●	◑	high
Liu [129]	◑	○	◑	○	●	high
Liu et al. [130]	◑	◑	◑	◑	●	high
Liu et al. [126]	◑	◑	◑	◑	●	high
Liu et al. [122]	●	◑	◑	◑	○	high
Liu et al. [125]	○	◑	●	●	◑	high
Liu et al. [123]	◑	◑	●	◑	●	high
Loi [133]	○	◑	●	●	○	high
Lu and Wu [134]	◑	○	◑	○	●	high
Lucca [135]	◑	●	◑	◑	○	some concerns
Lukassen et al. [136]	●	○	●	○	●	high
Lukauskas and Šarkauskaitė [137]	○	○	●	◑	●	high
Luo et al. [138]	◑	◑	◑	◑	○	some concerns
Lyu and Xia [139]	○	◑	◑	◑	●	high
Ma et al. [144]	○	○	◑	◑	○	some concerns
Ma et al. [140]	◑	◑	◑	●	◑	some concerns
Ma et al. [142]	○	◑	◑	◑	◑	some concerns
Ma et al. [141]	●	○	◑	◑	●	high
Ma et al. [143]	◑	◑	●	○	○	some concerns
Madrid-García and Merino-Barbancho [146]	○	○	◑	◑	◑	some concerns
Mazzoni and Failla [148]	◑	◑	◑	◑	●	high
Mei et al. [149]	◑	◑	◑	●	◑	some concerns
Messoudi and Mansar [150]	◑	○	●	◑	●	high
Min and Chen [153]	○	◑	◑	○	●	high
Mohebifar and Bit [155]	●	◑	◑	◑	○	high
Mu et al. [156]	●	○	◑	◑	○	high
Munot [158]	◑	○	◑	○	●	high
Nandakishore [159]	◑	◑	◑	◑	●	high
Nandakishore [160]	●	○	○	○	◑	high
Pan et al. [164]	○	◑	●	◑	○	some concerns
Panchal [165]	◑	◑	◑	◑	◑	some concerns
Paramasivam and Rajkumar [166]	◑	◑	●	◑	◑	some concerns
Paschalides et al. [167]	◑	○	◑	◑	●	high
Pei et al. [168]	○	◑	◑	◑	◑	some concerns

Study	D1	D2	D3	D4	D5	Overall
Pelaseyed and Zabihi [169]	◑	◑	◑	◑	◑	some concerns
Perifanis et al. [170]	○	○	●	○	◑	some concerns
Polignano et al. [173]	○	○	◑	●	◑	some concerns
Porcedda [174]	●	◑	●	●	○	high
Qian et al. [175]	◑	◑	◑	○	●	high
Qin et al. [177]	◑	○	◑	◑	◑	some concerns
Qin et al. [176]	◑	◑	◑	●	◑	some concerns
Rafe and Das [179]	◑	○	◑	◑	○	some concerns
Rafe and Das [180]	○	○	◑	○	○	low
Rafe and Das [181]	○	○	◑	○	○	low
Rahimi et al. [182]	●	◑	●	◑	○	high
Rajput et al. [183]	◑	○	◑	○	●	high
Rao and Callison-Burch [184]	◑	○	◑	●	●	high
Ren et al. [186]	●	○	◑	●	○	high
Rezvani et al. [187]	◑	◑	◑	○	●	high
Robitza [188]	◑	◑	◑	●	◑	some concerns
Sadoun et al. [190]	○	○	●	○	◑	some concerns
Saeed et al. [191]	◑	◑	◑	◑	●	high
Sakhawat and Raiyan [192]	○	○	◑	○	◑	some concerns
Santillana [193]	◑	○	◑	◑	◑	some concerns
Sautenkov et al. [194]	●	◑	◑	●	◑	high
Severo et al. [195]	◑	●	◑	○	●	high
Shankaranarayana et al. [196]	◑	○	○	○	◑	some concerns
Sharma and Lall [198]	●	○	◑	◑	●	high
Sharma et al. [197]	●	○	◑	○	○	high
Sheng et al. [199]	○	◑	●	○	◑	some concerns
Shi et al. [200]	◑	◑	◑	◑	●	high
Song et al. [201]	◑	○	◑	○	◑	some concerns
Song et al. [202]	◑	○	◑	◑	●	high
Song et al. [203]	◑	○	◑	◑	●	high
Sonthalia et al. [204]	◑	◑	◑	◑	◑	some concerns
Stambler [207]	●	◑	◑	◑	◑	high
Steiner et al. [208]	◑	◑	◑	●	◑	some concerns
Su and Huang [209]	○	○	◑	○	◑	some concerns
Sudjianto and Burakov [210]	◑	○	◑	◑	◑	some concerns
Sun et al. [211]	◑	◑	◑	◑	●	high
Sun et al. [212]	◑	◑	●	○	◑	some concerns
Tan et al. [213]	●	◑	◑	◑	◑	high
Tang and Chen [214]	◑	◑	◑	◑	●	high
Teke [215]	●	○	◑	●	◑	high
Teke [216]	●	○	◑	●	◑	high
Teng et al. [217]	◑	◑	◑	◑	●	high
Terven et al. [218]	◑	○	◑	○	◑	some concerns
Tommasino and Rinaldi [220]	○	○	◑	◑	◑	some concerns
Minh et al. [154]	◑	◑	◑	○	●	high
Vats [231]	○	○	◑	○	◑	some concerns
Wake and Wagle [233]	◑	◑	◑	◑	◑	some concerns
Wang and Zang [242]	●	○	◑	●	●	high
Wang et al. [236]	◑	●	◑	○	●	high

Study	D1	D2	D3	D4	D5	Overall
Wang et al. [241]	◑	◑	●	◑	◑	some concerns
Wang et al. [246]	◑	◑	●	◑	◑	some concerns
Wang et al. [238]	●	◑	●	◑	◑	high
Wang et al. [240]	◑	◑	●	◑	◑	some concerns
Wang and Gao [243]	◑	◑	●	◑	◑	some concerns
Wang et al. [237]	◑	◑	●	◑	●	high
Wang et al. [239]	○	◑	◑	◑	◑	some concerns
Wang [234]	◑	◑	◑	●	●	high
Wang et al. [245]	◑	◑	◑	◑	◑	some concerns
Wang et al. [244]	○	◑	◑	◑	●	high
Wang et al. [235]	◑	◑	●	◑	◑	some concerns
Weng et al. [247]	◑	○	●	●	◑	high
Wu et al. [252]	○	◑	◑	●	◑	some concerns
Wu and Lim [253]	●	○	◑	○	●	high
Wu et al. [251]	●	◑	○	●	○	high
Wu and Lim [254]	◑	◑	◑	◑	◑	some concerns
Wu [255]	◑	◑	◑	●	●	high
Xie et al. [256]	◑	◑	◑	○	●	high
Xu and Mei [258]	◑	◑	○	◑	◑	some concerns
Xu et al. [260]	◑	○	◑	○	●	high
Xu et al. [257]	◑	◑	◑	●	●	high
Xu et al. [259]	◑	◑	◑	◑	◑	some concerns
Yadav [261]	◑	○	●	●	●	high
Yamamoto and Kuroki [262]	◑	○	◑	●	●	high
Yan [263]	◑	◑	●	◑	●	high
Yan [264]	◑	◑	◑	●	●	high
Yang et al. [267]	◑	◑	◑	○	◑	some concerns
Yang et al. [266]	◑	◑	◑	◑	◑	some concerns
Yang et al. [268]	●	◑	●	○	○	high
Yang et al. [265]	◑	◑	◑	○	◑	some concerns
Ye et al. [269]	●	○	◑	◑	○	high
Yilmaz et al. [270]	◑	◑	●	◑	◑	some concerns
You et al. [271]	◑	◑	●	○	●	high
Yu and Yao [272]	◑	○	◑	○	●	high
Zhan et al. [273]	◑	◑	◑	◑	◑	some concerns
Zhang et al. [277]	◑	◑	◑	◑	●	high
Zhang et al. [280]	◑	◑	◑	◑	◑	some concerns
Zhang and Yu [281]	◑	◑	◑	◑	●	high
Zhang et al. [274]	◑	○	●	◑	◑	some concerns
Zhang et al. [278]	○	○	◑	●	◑	some concerns
Zhang [279]	○	◑	●	◑	●	high
Zhang et al. [276]	◑	○	◑	●	●	high
Zhang et al. [275]	◑	◑	○	◑	●	high
Zhao et al. [282]	◑	◑	◑	◑	◑	some concerns
Zheng and Chen [283]	◑	○	◑	◑	●	high
Zhong et al. [285]	●	○	●	○	○	high
Zhou et al. [286]	◑	◑	◑	◑	◑	some concerns
Zhou et al. [287]	○	◑	◑	○	●	high
Zhu et al. [289]	◑	◑	●	●	●	high

Study	D1	D2	D3	D4	D5	Overall
Zhu et al. [288]	●	○	●	◑	◑	high
Zhu et al. [290]	◑	◑	◑	●	●	high
Zhuang et al. [291]	◑	○	◑	◑	◑	some concerns
Zhuang et al. [292]	◑	◑	●	●	◑	high
Hume, Archer [77]	○	●	○	○	◑	some concerns
Bespoke Labs [14]	◑	◑	◑	◑	◑	some concerns
Palmer, Jared [163]	◑	◑	●	●	◑	high
Convai Innovations [39]	◑	◑	◑	◑	◑	some concerns
LocalLLaMA [132]	●	◑	◑	◑	○	high
Balian, Rob [9]	◑	◑	◑	●	◑	some concerns
TypeSafe AI [230]	●	●	●	◑	◑	high
Rauch [185]	◑	●	●	◑	◑	high

Table S3. The deciding reason for every domain rated high in Table S2, with its location in the study. D1 to D5 are the domains of Table S2.

Study	Domain	Reason
Ahmed [1]	D3	Clef-Flash is identified only by name and Hugging Face path with no revision hash or API version string, and no run dates are given for the zero-shot API runs (Section 4.1; Table 1).
	D4	The fine-tuned Clef-Flash checkpoint (epoch 4) was selected on test accuracy because no validation set was held out, while the TF-IDF baselines were tuned by cross-validation on training data only (Section 3.3).
	D5	Symptom2Disease is a public Kaggle dataset predating 2026 with labels released and no contamination check is reported (Section 3.1).
Amit [2]	D5	Open-Jev (2025) and the xLAM, Aegis 2.0, ToolACE, WildGuard and repo prompt-injection sources are public pre-2026 datasets with no contamination analysis (Sections IV, VI).
Antunes et al. [4]	D3	ndm-base is described only by its training stages (mmBERT-base through stages 0, 1 and 2a) with no release identifier, checkpoint link or run date (Section 10).
Asseri and Asseri [5]	D5	Every scored item is a verbatim VSM 2013 question, and the Hofstede reference scores it is compared with were public long before 2026, with no contamination check (Sec. 3.2; Sec. 3.3).
Azizi et al. [7]	D1	The headline suite GuardBench has ground truth computed from generating attributes, a synthetic procedure; public screening labels are undescribed (Sec. 5).
	D3	Open models (laya-td, laya-en, laya-ml, von, rlcd, Qwen2.5) are named without checkpoint revisions, hashes or run dates (Sec. 2).
	D4	The positive policy phrasing was chosen for the attack experiments after the wording turned out to matter, so that gates work before attack, a setting selected after seeing results (Sec. 5 GuardBench).
Bailo [8]	D1	Labels come from doomLaya's rule-based oracle, a synthetic procedure the author says a reasonable player can disagree with (Sec. 3, 6).
	D3	Evaluated open models are identified by name and size only; no revision hash or run date, and only the llama.cpp and doomLaya commits are pinned (Sec. 3).
Bellver-Soler et al. [11]	D3	ASOM has no version or checkpoint identifier (weights 'upon acceptance') and the audio LLM and CLAP comparators carry no revisions or run dates (Section 1 footnote; Section 4).

Study	Domain	Reason
	D4	ASOM's design choices were made by ablation on the same Dev-select partition that supplies the headline scores and is 'used for model selection', while the zero-shot audio LLMs receive no tuning (Section 4 Evaluation data; Section 5.4).
	D5	MSP-Podcast, CREMA-D, EmoBox corpora, EmoNet-Voice and EARS are public pre-2026 datasets used with their released labels, the headline partition doubles as model selection data, and no contamination check of the frozen encoders or audio LLMs is reported (Section 4).
Besanson [12]	D1	Error rates and verdict exposure are scored against the truth label of each seeded synthetic record, a value fixed by the generator before the document text is produced, so the reference is a synthetic procedure (Sec 6 Documents and labels; Sec 5). [G5]
Brom [16]	D5	MASSIVE and the four extension datasets are public pre-2026 sets used with their gold labels, no contamination check is reported for Jev, and only exact duplicates were removed from the Laya fine-tuning set (Sections 5, 9, 14).
Buecheler et al. [18]	D3	Jev is pinned (jev-1.13-20260917, 30 September), but the evaluated GPT-4o and GPT-4o mini were queried through rolling aliases rather than dated snapshots (Sec. 4.7; App.
Bukhari et al. [19]	D4	The hand-written 97-case set was consulted repeatedly and the v3 and v4 recipes were developed after inspecting earlier results on it, so it is not an independent final test, and the 'Paper' comparison columns come from a different training condition (Sections 5 and 10; Table 1 caption).
Cao et al. [21]	D5	In-distribution tests are converted items from public pre-2026 datasets (VitaminC, LCQMC, PAWS-X, FRAMES, BrowseComp-Plus) and no contamination check is reported for Jev or Qwen, only split isolation of the authors' own training data (Appendix; Section 7).
Chen and Li [23]	D1	The reference is the exact posterior computed by a generator for synthetic finite worlds, a synthetic procedure (Section 4.1; Section 5.1).
Chen et al. [27]	D2	No sample size, split, sampling design, interval, seed or repeated run is reported for any configuration, and the Conclusion's 93.54% on VoxCeleb1-O disagrees with 93.7595 in Table 3 and the abstract (Section 4; Table 3; Conclusion).
	D3	The evaluated checkpoint is described only as an MoE decoder, with no checkpoint name, version, release date or run date, and the text does not say whether SFT or RLCD weights were scored (Sections 2, 3, 4).
	D5	All 18 evaluation sets are public pre-2026 benchmarks used with their released labels, the model is jointly trained on heterogeneous audio datasets including MELD and VoxCeleb, and neither the per-set split nor overlap with training is stated or checked (Section 1; Section 4.1).
Chen [25]	D3	Table 5 lists compared endpoints such as cloudflare/clef, liquid/d1 and inception/mercury-decide with no version or snapshot, and no run dates are given (Table 5; App.
Chen et al. [26]	D3	The evaluated models are the officially released AID weights and the authors' UCM LoRA checkpoint with no version identifier, hash or run date (Section III-A; Section IV-C).
	D5	UCM, AID and NWPU-RESISC45 are public benchmarks predating 2026 used verbatim with labels and no contamination check is reported (Section IV-A).
Cheng et al. [31]	D3	MetaEncoder, OpenJev-27B and the embedding baselines are named with a repository link only, no revision and no run date (Sec. 4.3; refs).
	D5	JevBench, ImaJev, MMLU, MMMU and the other suites are public pre-2026 with answer keys and no contamination check; MetaEncoder training data are not described (Sec. 4.1).
Cheng et al. [32]	D1	the benchmark labels are generated by a simulator, a synthetic procedure.

Study	Domain	Reason
	D3	No Jev version string anywhere; Jev replies were recorded by NanoJev or run by an unnamed third party; own model named only as 1.0 (note Decision models).
	D4	this-that-model-1.0 trained on all 15 families of the benchmark it is scored on while every comparator is zero-shot; the authors call it a weak scientific comparison (note Test-set exposure; fulltext 4.3).
Chouhan and Perry [33]	D3	Jev is called through the unpinned gateway alias typesafe-ai/jev that 'does not expose a pinned semantic version', so only run dates (2026-09-25 to 27) and the vendor card's Jev 1.13.0 figure stand in for a version (Appendix L; Limitations).
	D4	The five-seed large-model run was launched only if base-model test accuracy cleared a margin, and live test accuracy was seen during training (Limitations, Selection and testing).
Chung et al. [35]	D5	The nine benchmarks are public releases from 2019 to 2026 used with their keys, and the authors state training exposure cannot be excluded with no contamination check (Limitations; Section 3.2).
Cinà [36]	D3	Jev is only 'jev-latest, September 2026' with no version string or run date, and the Laya checkpoint is 'open 421M' with no revision (Section 4 Arms; Section 5.4).
	D4	The 0.2-point matching bound behind the headline 'matches' claim was fixed after seeing the test results, a setting selected after seeing results (Sec. 4 Statistics; Sec. 5.4).
Dai et al. [40]	D3	Jev has no version identifier and no run date in the paper (Section 4.1, Table 2).
Dai et al. [41]	D1	Reference probabilities are exact analytic class probabilities from simulated text worlds, a synthetic procedure (Section 3; Appendix B).
	D3	Only Jev carries a dated snapshot (typesafe/jev-1.13-20260917), while compared models Clef and Solar Decide are listed by bare product name with no release token (App.
Denney and DiGiuseppe [45]	D5	The recall tasks (V-Dem v14, CHES, CMP, TEV) and the tweet and GTD sets are public and predate the models, the authors say agreement may partly reflect recall, and no contamination check is run (Sections Party positions and Democracy indicators; note Test-data exposure).
Ding et al. [47]	D4	The headline DecisionBench and typed-decisions Jev figures are third-party leaderboard values, and Bongard is a trained system with temperatures fitted on external development sets while Jev is zero-shot (Table 8; Sec. 6, 7.1).
Ding [48]	D5	MMLU (2020) is a verbatim public benchmark with released answer keys and the authors state their checks 'cannot rule out paraphrases or pretraining contamination' (Section 6.1).
Dinzinger et al. [49]	D5	TREC DL and BEIR are public benchmarks with qrels released before 2026, MS MARCO exposure of the reranker fine-tunes is acknowledged, and no contamination check of the Qwen backbones is made beyond a zero-shot subset defined by fine-tuning data (Section 4; Section 7).
Dong et al. [50]	D1	Cases and decisions are authored by the researchers (synthetic) and the transfer-set labels come from blinded AI review, not human adjudication (Sec. V-A).
	D4	The question clean control and the corrected full-trace audit were specified after defects were found on the same cases, so their settings were selected after seeing results (Sec.
Rosario and Atef [189]	D4	Every prompt, threshold, grid and cascade parameter was chosen on the same 495-claim development split that is scored, with outcomes visible (Sec. 5; Sec. 12).
	D5	LLM-AggreFact is a public benchmark released before 2026 and the authors state neither system can be assumed uncontaminated, with no contamination check (Sections 5, 12).

Study	Domain	Reason
dos Santos Barbosa [51]	D2	One run per arm against a single target, no intervals anywhere, and the paper states this itself: the results are observations that motivate the architecture, not statistically powered experimental evidence (IX-A).
	D3	The case study requests jev-latest with no patch-level build string, snapshot hash or run date recorded anywhere in the paper (Table V).
	D4	The reported re-test uses salvage and classification fixes chosen after failures on the same target (Sections IX-C, IX-E and IX-H, Table VIII. The earlier matched setup does not remove this C2 trigger).
Du et al. [52]	D5	Sources are mostly public pre-2026 datasets (MedQA, MIMIC, VitalDB, CTG, MEDEC) used with their answer keys and no contamination check is reported (Section 3; Appendix B).
Duan et al. [53]	D3	Jev is given no version identifier in the main runs (Table 9 shows blank settings) and 'Jev 1.13' appears only for Sections 5.4.1 and 5.4.2, with no run dates for any model (Table 9; Appendix C.3).
	D5	JEVal is resampled verbatim from 36 public datasets whose answer keys predate 2026, and no contamination check on the evaluated models is reported (Section 3.2; Appendix B).
Fadli et al. [56]	D5	NusaX (public since 2022) and NusaTranslation are verbatim public test sets with released labels; the authors state closed models may have seen them and no check is possible (Sec. 4.10).
Feng et al. [57]	D5	LaMP-2 and MIND are public pre-2026 datasets with released labels, and the paper does not discuss pretraining exposure of the Qwen backbone (Appendix A.3; Appendix A.4).
Gan et al. [60]	D5	The suite is built from public pre-2026 datasets with released gold labels, the authors cannot exclude backbone exposure, and only train-test leakage was checked with no contamination check of the pretrained backbones (Limitations; Appendix A).
Gao et al. [61]	D5	All 40 datasets are public pre-2026 sources with released labels and the authors state they 'did not audit item-level overlap' (Appendix B.2; Limitations).
Geuter et al. [63]	D1	The headline OpenDecisions labels come from LLM authoring agents kept only if a panel of three open models agreed, plus code-computed items and panel-answered real requests, with no human labelling (Section 3 Labels, Table 1).
	D4	The authors' training recipe was selected on OpenDecisions, the very set that carries the headline ranking, and each public system gets the better of its original or recalibrated output (Section 5 Setup, Section 4).
Guo et al. [66]	D5	Who&When is a public 2025 benchmark used with its annotations, Who&When Pro is public, and no contamination check is reported (Section 4.1).
Guo et al. [65]	D4	The text uses the in-sample best generic readout and selects a shared question on both evaluated states (Sections 3.4 and 4.2, Figure 4).
Ha et al. [67]	D4	For the perturbation results, each model uses the label-description configuration with the best mean macro-F1 on the original test data (Sec. 5.2; Table 4).
	D5	All 12 datasets are public corpora with released gold test labels, 11 of them from before 2026 (ViANLI is cited to a 2026 journal article), and the paper reports no contamination check (Section 3.1; Table 9).
Han et al. [68]	D1	The reference is the logged next hop chosen by the production LLM agent fleet, a model-generated label (Abstract; Sec. 5).
Hu et al. [72]	D1	Of 862 reference labels, 586 are Jev's own stable answers from ten repeated runs, 41 are a three-LLM consensus and only 185 were human reviewed (Section 3.1; Fig. 2b), so the target model largely labels itself.

Study	Domain	Reason
Huang et al. [75]	D3	Mercury Decide, a hosted endpoint, is evaluated with no version identifier, Jev carries only the minor version 1.13, and no run dates are given (Sec. 5.1; Table 2).
	D5	Benchmarks are public pre-2026 sets with answer keys and no contamination check, and the authors note that Torchcast 12B and Kev 9B were trained on BANKING77, which contains Banking20 (Section 5.1, Appendix D).
Huo et al. [78]	D1	Reference actions are sampled programmatically and every request is an LLM-written description with no human validation, a synthetic procedure (Sec. 2.3).
	D3	The comparator native JEV has no version, access date or API detail, and the flagship LLMs carry no revisions or run dates (Sec. 3.2.1, Table 3).
Ibrahim and Zaki [80]	D3	The Table 5 identifies the comparator as gemini-3.1-pro-preview without resolving the preview tag further (Appendix A.5, Table 5).
	D4	The paper selects the per-task best LLM on the same test items (Section 4.1).
	D5	All 18 tasks are public benchmarks that predate the models; the authors state contamination cannot be excluded (5, Discussion).
Jiang et al. [84]	D1	Correctness is scored by a gpt-4o-mini LLM-as-a-judge against LoCoMo answers (Table 1 caption).
	D2	Item count not stated, single run, no seeds or intervals; the setup says two benchmarks but only LoCoMo is reported (4.1; note n).
	D3	No Jev version string and no dates anywhere (note Decision models; Appendix B).
	D5	LoCoMo is a public 2024 benchmark with released answers; exposure not discussed (note Test-set exposure).
Joy [85]	D5	TREC, CLINC150 and MASSIVE are verbatim public benchmarks released before 2026 with training exposure 'unknown' and no contamination check (Section 8).
Khatri et al. [87]	D3	Jev 1.13.0 and 1.14.0 are pinned and the census date is given, but the authors state several evaluated identifiers are unversioned, e.g. d1 (Sec. 5; Sec. 7 Limitations iii).
	D5	Conflicts (2025) is a public benchmark with released labels, no contamination check is reported, and identical requests changed up to 9.4 percent of predictions (Section 3; Section 7).
Kourtis and Xilouris [91]	D1	All labels 'come from simulator rules' in a synthetic 5G testbed corpus, a synthetic procedure rather than annotation (Section V-B; Section IX item 5).
Lai et al. [94]	D1	All items, rules and answers are synthetic, with a programmatic oracle returning every label from executable arithmetic rules (Section 3; Appendix A).
Lan et al. [95]	D3	Jev is queried as 'jev-latest' with no resolved version and no run dates, only a blog access date of 2026-10-01 (Section 3.2; references).
	D5	All four benchmarks are public and mostly pre-2026 and used verbatim with their labels, and the paper does not test contamination of the evaluated models (Section 3.1).
Lee et al. [97]	D5	MedQA, USMLE, DiagnosisArena and the NEJM vignettes are public pre-2026 items with released keys and the authors say pretraining overlap cannot be excluded with no contamination check, while overlap of the MIMIC-IV Note adaptation text with the MIMIC test patients is not discussed (Section 5).
Lee and Park [98]	D4	The best of 28 pyKT baselines is chosen on test in every cell; it favours the comparator, but it is still selection on test data (Table 1 note; Sec. 4.1).
Li [111]	D5	CMMLU, C-Eval, TMMLU+, SeedBench, CROP and Ag-valuate are public pre-2026 benchmarks used with answer keys; exact-overlap exclusion covers only the authors' training data and backbone pretraining exposure is unresolved (Section 4.1; Appendix A).

Study	Domain	Reason
Li et al. [114]	D5	VAST (2020) and ZS-CSD (2025) are public test sets with released labels and no contamination check (Sec. 4.1).
Li and Xie [108]	D4	PPLX and Jina were chosen after exploratory outcomes, the input-budget settings and margin were set after seeing BRIGHT results while Jina ran at default (a tuned Jina is only a control), and a higher-scoring Winnow-12B was dropped without documented reason (Section III-B; Section IV-C; Section V).
Li et al. [109]	D3	Commercial Jev has no version or run date, and its initialization source is 'Not publicly confirmed' (Table 3).
Li et al. [101]	D4	the comparator interface was changed and chosen after seeing the first evaluation results.
Li [105]	D2	Counts do not reconcile, since points A to F sum to 3,620 (RAG relevance n = 1,920) against a stated 2,100-case suite, and no tally matches 7,283 base cases (Abstract; Sec. 4; Table 1).
	D5	All 18 sources are public pre-2026 and the author says near-perfect scores on popular sets 'may partly reflect exposure', with new controls on only three decision points (Section 8).
Li and Wang [107]	D5	BANKING77, MASSIVE and Mind2Web are public 2020-2023 benchmarks used with their labels, the curriculum includes similar intent-selection data and no contamination check is reported (Section 4; Appendix C.1).
Li et al. [116]	D4	The text adds model families after earlier results were known and evaluates them on the same 1,312 items (Section 3, What was frozen and when).
Li and Chen [110]	D4	The headline 95.4% uses curated avoidance menus on the 500 CLINC150 test items, built with the gold label offline, against a 4B model on the full menu (Sec. 5.3; Sec. 6).
	D5	CLINC150, MMLU, ARC and the other sets are public benchmarks from 2018 to 2021 with released answers, and no contamination check is reported (Appendix A; note Exposure).
Li [104]	D5	public pre-2026 datasets where contamination could not be ruled out.
Li and Wagle [117]	D4	Training mixtures were iterated using error analysis on public JevBench samples, items later reported as test results, and the baselines are finished systems with their own training (Section 3.5).
Li et al. [115]	D1	Gold actions and probe answers are stored labels computed by program from simulator state or databases (shortest plans checked by breadth-first search) and matched against model choices (Sec. 3).
	D3	OpenAI Decisions is evaluated as an unversioned public-beta endpoint with no version identifier, even though Jev and the open-weight systems carry release names or pinned revisions (Table 1; Sec. 4).
Li et al. [113]	D4	The Prompt Guard 2 threshold is set at the 99th percentile of its scores on the same clean cases used for reporting, a setting tuned on test data (App.
	D5	Cases are verbatim public benchmark items (JailbreakBench 2024, InjecAgent 2024, a public phishing corpus) and the cookbook battery is used as published, with no contamination check (Section 4.2).
Li [103]	D4	The released Iris numbers use a checkpoint and an abstention margin chosen on the cross-task test split, while the retrained OneJev baseline is reported at its final checkpoint with a different LoRA rank (r=64 versus r=32) (Section 5.4; Table 2; Appendix A).
	D5	Test steps are verbatim public Multimodal-Mind2Web and MMEB-V2 items with public gold labels from before 2026, no contamination check of the Qwen3.5 backbone is made, and the cross-task split also served for selection (Section 4; Section 5.4).

Study	Domain	Reason
Li et al. [100]	D1	Correctness on the constructed access-policy, placement, contract-catalogue and route tasks is the minimum-cost feasible selection or justified refresh or escalation computed by complete feasibility and cost predicates, a program-made reference compared with each model's answer (Sec IX; Tables IX, XI, XII captions). [G5]
	D3	Jev appears only as the minor version Jev-1.13 on the authors' own tasks with no run date, and the MiniLM re-ranker compared on the contract task is a bare name with no checkpoint, so one compared system has no version identifier (Sec VII model-version paragraph; Table X; Table VII caption). [G2]
Li et al. [112]	D3	No Jev version string or run date appears; only source access dates (30 Sep 2026) are given (Sec. 3.1; App. C).
	D5	All 13 benchmarks are verbatim public pre-2026 sets with released answer keys and no contamination check (Sec. 3.1, 3.2).
Li et al. [99]	D1	Verifier labels come from GPT-5.4 annotation with unreported human verification, and the downstream score is a GPT-5.4 composite (Sec. 3.4, 4.1, 4.3, 7).
Lin [119]	D1	The 324-item frozen holdout is 'synthetic, model-labelled' with labels from the same family of models as the training labels (Section VIII; Section V-A).
Liu [127]	D1	Candidates are authored mutations of DROP and GSM8K answers whose required verdict follows from a construction certificate, a synthetic procedure with no human adjudication (Sec. 3; Eq. 1).
Liu et al. [124]	D5	All three datasets are public pre-2026 logs (Goodreads, HetRec LastFM 2011, Amazon Reviews 2023) with no contamination check, and the authors say cold items may have appeared in LLM pretraining (Section 4.1; Appendix A.2).
Liu et al. [128]	D3	Only Jev (1.13.0) and pplx (v1.1) carry versions, while d1, clef, GPT-6 Luna and the gpt-6-sol reference have only unversioned aliases marked not reported, and no run dates are given (Appendix B, Table 5).
	D4	Cascade members, order and deferral rates are chosen on the same evaluated items, and five-fold cross-fitting covers only the rates (Sec. 3.4; Limitations).
Liu [129]	D5	R-Judge (2024) and AgentHarm (2025) are verbatim public benchmarks with released labels, and the paper says they 'may overlap training data' without a contamination check (Section 8).
Liu et al. [130]	D5	MMEB-V2 (VLM2Vec-V2, 2025) is a public benchmark used verbatim with released relevance annotations and no contamination check, and ViDoRe V3 (2026) is public too, so the high rating rests on MMEB-V2 (Section 4, references 22 and 24).
Liu et al. [126]	D5	All 13 sources and Dreaddit are public pre-2026 datasets used with their released labels, zero-shot Jev and the Qwen backbones may have seen them, and no contamination check is reported (Section 3; Section 5.7).
Liu et al. [122]	D1	Decision labels come from an LLM extracting rules and generating synthetic patients, cross-verified by a different model family with two further families for disputes, and grading labels are code-computed from sampled parameters, with no clinician adjudication (Section 3.1; Section 6).
Liu et al. [125]	D3	The Jev model identifier and version are not reported (only 'Jev AI (2026) developer documentation') and no run dates are given (Section 4.5; References).
	D4	'The gate configuration was selected on these layouts, making this a development-set evaluation', so the Jev condition was tuned on the same layouts used for the comparison (Section 4.5; Section 5).
Liu et al. [123]	D3	Decider-4B and Decider-2B are named with no version, revision or checkpoint hash and no run dates (Section 5.1, Table 1).
	D5	TSQA is a public 2025 benchmark whose test items and answers predate 2026, with no contamination check (Section 5.1).
Loi [133]	D3	The returned alias does not identify a verified served version (App).

Study	Domain	Reason
	D4	Topics were chosen partly because earlier explorations made them promising, 16 repetitions were adopted after reviewing earlier results, and two ratings were admitted retrospectively (Sec. 3; App.
Lu and Wu [134]	D5	All five sources are public pre-2026 datasets used verbatim with labels, the paper states pretraining overlap is unknown and no contamination test is run (Section 4.1; Limitations).
Lucca [135]	D2	The relationship column of Table 1 sums to 239 against the stated corpus of 231 emails (the type and urgency columns sum to 231), so the reported counts do not reconcile (Table 1; Section 5).
Lukassen et al. [136]	D1	Better and worse responses are fixed by construction, with GPT-4-Turbo writing both fact-related responses and alignment pairs taken from preference datasets, a synthetic procedure (Sec. 3.1 Data).
	D3	The compared TDM Clef is returned only as the bare string cloudflare/clef with no version or date, while timing is given only as early October 2026 (Sec. 3.2; Table 6).
	D5	All items are public CALM items (released 2025) from older public datasets, and the authors state the models may have seen some of them, with no contamination check (Limitations).
Lukauskas and Šarkauskaitė [137]	D3	Most decision models resolve to dated versions, but Clef and Clef Flash resolve only to cloudflare/clef and cloudflare/clef-flash with no version (Table 13; Sec. 4).
	D5	The human survey data have been public since 2020 and the authors state they did not test the decision models for memorisation; the C0 argument is indirect, not a contamination check (Section 7 Memorisation).
Lyu and Xia [139]	D5	Amazon Reviews 2023 leave-last-out splits, whose interactions predate 2026, are used with the public held-out item as the answer; the authors note possible familiarity but run no contamination check (Sec. 4.1; Sec. 6.2).
Ma et al. [140]	D4	Larger pruning budgets were selected as the largest budget keeping 90% retention using scores on 128 questions per public set and all 188 LIBERO frames, which are test data; only the six-block default was fixed in advance (Sec. 4.6, App. B.2).
Ma et al. [141]	D1	Wrong-value labels on CFB and FinTrace come from the authors' rule-based construction and FCRB labels from the BFCL checker program, with only a 100-item human audit (Sec. 3.2; Appendix A).
	D5	FCRB and ComplexFuncBench are public 2025 benchmarks with released labels and FinTrace is a public April 2026 resource; no contamination check is reported for Jev (Section 3.2).
Ma et al. [143]	D3	Jev is named only through 'Decision model documentation' accessed 2026-10-02 and no model version, API access or run date is reported (Section 1; reference 33).
Mazzoni and Failla [148]	D5	PluRule, OLID, HateXplain, HateModerate and ModerationBench are public datasets with released labels from before 2026 and no contamination analysis is reported (Section 3.1; note Exposure).
Mei et al. [149]	D4	MC11 is the evaluation benchmark, and the one-epoch lower-rate recipe used for every compared model was chosen from MC11 training-recipe diagnostics (Sec. 5.4; Table 4).
Messoudi and Mansar [150]	D3	Evaluated endpoints upstage/solar-decide and respan/span-01 carry no version or snapshot, Jev is typesafe/jev-1.13 only, and the run window is approximate (App.
	D5	Five of six datasets are public pre-2026 benchmarks used with released labels and the authors state they cannot rule out contamination for API models, with no check run (Ethical Considerations; Table 1).

Study	Domain	Reason
Min and Chen [153]	D5	Items are the verbatim AppWorld test-normal split and OfficeBench tasks, both public since 2024 with evaluation checks, and no contamination check is reported (Sec. 5.1).
Mohebifar and Bit [155]	D1	References are synthetic, generated by GPT-6-Sol, and only 150 of 1,968 cases were engineer-reviewed with no final adjudicated key (Section III; Section III-A).
Mu et al. [156]	D1	Labels come from the simulator's event generator in stochastic EPANET runs, and one hydraulic model generates both events and the digital twin (Sections 2.1, 4.4).
Munot [158]	D5	MetaTool (2024), HotpotQA (2018) and LexGLUE ECtHR (2022) items are used verbatim with public answer keys, and no contamination check is reported (Section 2 Datasets; Limitations).
Nandakishore [159]	D5	RewardBench, RewardBench 2, RM-Bench, JudgeBench and HaluEval are public 2024 and 2025 releases with answer keys and no contamination check is reported (Section 3; note Exposure).
Nandakishore [160]	D1	Gold is the mean of three samples from a roughly 4B-class teacher endpoint and 'every number in this paper measures teacher agreement on synthetic cases' (Section 3.2).
Pan et al. [164]	D3	Jev and every other decision model are named without any version or revision identifier and the only dates are documentation access dates of 13 October 2026, with no run dates (Table 1; Appendix B.3-B.4; references).
Paramasivam and Rajkumar [166]	D3	Hosted Jev is cited only by a documentation URL with no version string or access date, and the RubriX checkpoint has no hash (Section 4.3; reference 25).
Paschalides et al. [167]	D5	Dynamic, MHS, HateXplain and HateCheck are public 2019 to 2021 datasets used on their test partitions, and the authors state prior training exposure remains uncertain with no contamination check (Section 5.1).
Perifanis et al. [170]	D3	Jev's version string is not reported anywhere in the text, Laya's version is not stated, and run dates are not given (Section 4.12; References).
Polignano et al. [173]	D4	All design choices were made on the same development users that are reported, baselines are the authors' own implementations, and the Jev comparison is zero-shot versus Kev trained on in-domain records, with fresh-user confirmation not yet scored (Section 6.3, Section 9.2).
Porcedda [174]	D1	Reference probabilities are exact by construction from templated problems, a synthetic procedure (Section 2.1).
	D3	No Jev version identifier and no run date appear anywhere in the paper.
	D4	The inverse uncertainty map is fitted on the same 365 items it is scored on, with no held-out split, while SemIf gets no comparable adjustment and its setup is not described (Sections 3.1, 4, 5).
Qian et al. [175]	D5	GSM8K, MATH-500, AIME, AMC, GPQA-Diamond, SUM and UMWP are public pre-2026 benchmarks with answer keys and no contamination check (Section 4.1).
Qin et al. [176]	D4	The model is trained on 12.3M QA drawn from SAT, RefSpatial and VST-P sources that overlap the evaluation benchmarks while generalist baselines are not, and the token cap differs between early (8,192) and later (32,768) runs (Sec. 4.1; Appendix D).
Rahimi et al. [182]	D1	References are exhaustive-search optima and a utility from the authors' own simplified numerical simulator, a synthetic procedure on simulated channels and intents (Section IV-B, IV-C).
	D3	No Jev version, model ID or run date is given, the comparators are named only 'GPT-Sol' and 'Qwen' without identifier, and only a blog access date of 23 September 2026 appears (Section IV-A; References).
Rajput et al. [183]	D5	BANKING77, CLINC150 and AG News are public benchmarks predating 2026 used with their labels and no contamination check is reported (Section 3.2).

Study	Domain	Reason
Rao and Callison-Burch [184]	D4	The 5-point parity margin behind the headline replacement claim is post hoc, and cascade gains are also reported at oracle thresholds chosen on the evaluated pairs (Sec. 3.4; Sec. 6; Table 3).
	D5	All nine panels come from public benchmarks that predate 2026 and the authors say a judge may have seen the items or labels, with no contamination check (Limitations).
Ren et al. [186]	D1	Every test prefix receives its scenario ground truth (Section 4 Data, Appendix B. G5 gives high).
	D4	The recipe was chosen from about ten variants scored on the test set ('selected with test-set exposure', Abstract; 4 Model selection) while the judge prompt was untuned.
Rezvani et al. [187]	D5	CyberSecEval Instruct (Meta, 2023) is a public pre-2026 benchmark with released CWE labels and no contamination check is reported (Section III-A1).
Robitza [188]	D4	Input states were 'iteratively improved' over five versions in the author's own tests, and the headline uses S4, while comparators were trained with leave-one-source-out CV or are fixed standards (III-D, III-E).
Sadoun et al. [190]	D3	No Jev version string is given and only a documentation access date (2026-09-26) appears in the reference list (Appendix M; reference list).
Saeed et al. [191]	D5	All data are public pre-2026 Hugging Face benchmarks (AG News, SST-2, BoolQ, ARC, Banking77, CLINC150, MMLU) with no contamination check (Section 4; Data and Code Availability).
Sautenkov et al. [194]	D1	The reference is the known location of a synthetic incident composited into satellite maps, a synthetic procedure, scored by a footprint-overlap rule (Section 4 Scenarios and maps; Measures).
	D4	The only decision system's questions and thresholds were chosen on the benchmark maps used for reporting with no held-out set, and collision and oil-spill questions were refined after the first runs and the Gemini missions repeated (Section 4 Protocol; Section 6).
Severo et al. [195]	D2	The abstract and highlights report a 300-flow pilot with 5,400 decisions and GPT-5.6 Luna (F1 0.859), but the body reports 2,000 flows, 6,000 decisions per finite k and Gemini 3.6 Flash (F1 0.856), so the headline counts do not reconcile (Abstract, Section 5.3, Table 2).
	D5	NSL-KDD (2009) is a public benchmark with released labels and no contamination check is reported (Section 5.3).
Sharma and Lall [198]	D1	Accuracy is graded CORRECT/WRONG by a gpt-4o-mini LLM judge, and the human sensitivity grading covers only the 141 judge-discordant questions and was done by the first author (Section 4 Stack; Section 5.4).
	D5	LoCoMo (2024) and LongMemEval are verbatim public benchmarks with released answers, and the 'held-out' conversations were only unrun by earlier systems, with answer-model exposure unaddressed (Section 4).
Sharma et al. [197]	D1	Labels are computed by executing a stated rule over declared facts in programmatic generators, a synthetic procedure (Sec. 3.2).
Sheng et al. [199]	D3	Jev is 'accessed through its official API' with no version identifier and no run date, only a vendor-blog access date of 2026-09-26 (Section 5.1, reference list).
Shi et al. [200]	D5	iNews (ACL 2025) test items are used verbatim with released ratings, and no contamination check is reported for any evaluated model (Sec. 3; App.
Song et al. [202]	D5	ReviewCritique (2024) and ReviewScore (2025) are public datasets used verbatim with released labels over papers that may be in model training data, and no contamination check is reported (Section 3.2; Ethical Considerations).
Song et al. [203]	D5	All four datasets are public pre-2026 benchmarks used verbatim with released answer keys and no contamination check is reported, although the design targets presentation sensitivity rather than absolute accuracy (Section 2.1).

Study	Domain	Reason
Stambler [207]	D1	Gold solutions for the 128 puzzles and 47 transfer puzzles are produced by a generator (synthetic), and the form labels were written by the author except 41 SpreadsheetBench-derived fills, not independently validated (Sections 4.1 and 7; Appendix A).
Steiner et al. [208]	D4	The prompt F0 for each model and benchmark is the one with the highest test-set F1 and is reused in the calibration, cascade, cost and published-results comparisons (Sec. 4; Table 6).
Sun et al. [211]	D5	BrowseComp-Plus and NeuCLIRBench are public benchmarks released in 2025 with answer keys and no contamination analysis (Section 4.1).
Sun et al. [212]	D3	No version string for any of the three models in the text; checkpoints are deferred to the harness and the hosted model was reached 'at one point in time' (footnote 2; 7).
Tan et al. [213]	D1	Attack success is scored by two model-based evaluators, the HarmBench classifier and the fine-tuned StrongREJECT evaluator, with no human adjudication of the outputs (Section 6.1; Appendix A.5).
Tang and Chen [214]	D5	Test states are verbatim images and labels from public pre-2026 datasets such as MedMNIST, only re-split by the authors, and no backbone contamination check is reported (Sec. 4.1; Appendix A).
Teke [215]	D1	Gold is mostly AI-made (ai_ensemble_unanimous 1,659 and ai_rubric_agreed 1,157 of 4,275 questions), with only 252 owner-blind, 181 court-record and a few other human labels (Section 3.1, Table 2).
	D4	The author's own ufakzeka-karar had data shaped by reading test results in guardrails, moderation and customer support, so a compared system was tuned on test items (Sections 6, 7).
Teke [216]	D1	Most HakemBench gold was settled by passes of AI models and only 260 questions carry human-decided gold, while the support development set and human check are answered by one person, the author (Section 6; fulltext Section 8).
	D4	Run 3 was trained on data aimed at Run 1's errors on the full test set and its protocol was fixed after reading Run 2's test results, while comparators are zero-shot, and the author flags the numbers as 'shaped by reading the test results' (Sections 4 and 6).
Teng et al. [217]	D5	Banking20 and Newsgroups test items come verbatim from public pre-2026 datasets (2020, 1995) with labels, and no contamination check is reported (Sec. 5.1).
Minh et al. [154]	D5	Dataset B, BUSI and BUS-BRA are public pre-2026 image sets with released labels, and the authors state the evaluation cannot determine training exposure (Sec. 5.4).
Wang and Zang [242]	D1	Scoring uses BBH and MuSR answer keys from templated or generated reasoning tasks (coded synthetic in the ledger; the paper does not describe provenance) plus a human-designed object-placement workflow (Section 4; Appendix H).
	D4	The three best of six BBH tasks were chosen after inspecting broader results including held-out results, the truth-chain program was refined by agreement on ten development inputs, and the task-conditioned envelope is post hoc, so the authors say estimates describe selected task families (Appendices C and H).
	D5	BBH (2022 to 2023) and MuSR (2024) are public benchmarks with released answers and no contamination check is run for the teacher or Qwen, and development inputs were reused in admission checks (Appendix H).
Wang et al. [236]	D2	Per-dataset sample counts, splits, window lengths and call counts are not reported, no interval or repeat accompanies any score, and the exclusion of 14 archive datasets as 'datasets with available scores' is unexplained (Section 4.1; Table 1).

Study	Domain	Reason
	D5	UCR, UEA, EasyTPP, ETT and stock series are public archives from before 2026 with released labels, and no contamination discussion is given (Section 4.1; references).
Wang et al. [241]	D3	CalOn, the evaluated model, is named without any version, release tag or weights link (printed as CalOn in Table 4), and the InstructUIE baseline is the bare name ZWK/InstructUIE, so compared systems carry no version identifier (Table 4; Table 10). [G2]
Wang et al. [246]	D3	Hosted Jev has no version identifier anywhere, only a reference accessed 2026-09-27 (Sec. 4.1; reference list).
Wang et al. [238]	D1	Ground-truth decisions come from GPT-4o-generated application states built so that truth differs from the attacker target, and the model-extraction reference is Jev's own answers (Sections III-A3, IV-A3; Appendix C).
	D3	No Jev version string, model ID or query date is given for the official API, only 'as of September 2026' for a download count (Section IV-B; Section I).
Wang et al. [240]	D3	JEV is named with no version string or access date in the text read (Sec. 4.2, 5.1).
Wang and Gao [243]	D3	The JEV version identifier is not reported and the four generative judges are named with no revision, temperature or run dates (Section 3.2; Section 4.2; Appendix C).
Wang et al. [237]	D3	Jev is cited only as 'TypeSafe AI, 2026' with no version string, provider or run date; the only date is a documentation access on 2026-10-12 (Section 5.1; references).
	D5	MATH, MBPP+, API-Bank and IDRBench I3 are public pre-2026 benchmarks used verbatim with released tests and answers, and no contamination check is reported (Section 5.1; Appendix C.3).
Wang [234]	D4	Mnemon was developed on LoCoMo and LongMemEval-S and consolidation was designed after seeing HaluMem and BEAM-100K results, while the 14 compared systems enter as published numbers under a different grader (Sections 6.1, 7).
	D5	All benchmarks are public pre-2026, 'none of the benchmarks is held out', and no contamination check is reported (Section 7).
Wang et al. [244]	D5	GPQA-Diamond, SuperGPQA, MedQA, MedXpertQA and MMLU-Pro are public pre-2026 benchmarks with released keys, and the authors state exposure and label errors remain possible with no contamination check (Limitations).
Wang et al. [235]	D3	The hosted TypeSafe Jev has no version reported, UniJev has no release identifier, and selected revisions for released baselines are recorded only in a registry outside the text (Table 2; Appendix A).
Weng et al. [247]	D3	JevBench commit and Jev 1.13.0 are given, but the compared DeepSeek Flash, a hosted endpoint, carries no version identifier and no run dates are reported (Table 3; Sec. 6).
	D4	The 35B was chosen when two builds' public scores (79.3 versus 80.2) were known, the 1B and drill arm were designed after seeing public-item failures, and 15 of 52 comparators were selected by the authors (Section 9, Selection).
Wu et al. [252]	D4	alpha = 0.5 and T = 10 are stated with no separate selection data, and alpha = 0.5 is the row of the Table III sweep that the text says identifies a 'sweet spot', a sweep run on the same 100 scored episodes per dataset that every ablation reports, drawn from the evaluation benchmarks, while the replanning interval of 25 steps is 'empirically' chosen on unstated data and Table I does not say whether the baselines were re-run (Sec IV-A; Sec IV-C; Sec IV-C.2, Table III; Sec III-E; Sec IV-B). [C2]
Wu and Lim [253]	D1	Safe and attacker-target actions are fixed in advance by the InjecAgent construction and the model's single selection is compared with them, a precomputed synthetic label (Sec. 3.1; Sec. 3.2).

Study	Domain	Reason
	D5	The reconstruction keeps InjecAgent (2024) user goals, attacker goals and attack text verbatim with known attacker-target tools, and no contamination check is reported (Sec. 3.1).
Wu et al. [251]	D1	Gold actions are generated offline by fixed policy rules in two synthetic workflows and approved by the project owner, a synthetic procedure with no blinded human review (Section 4.1).
	D4	Coverage-matched probability thresholds were set from the evaluation clean scores and the scope-aware repair was designed after seeing the original gate fail on the same tasks, so the compared gates were selected after seeing results (Section 4.3; Section 5.6; Appendix A).
Wu [255]	D4	Baseline figures are copied from Wu et al. (2026) and Sun et al. (2026) and not rerun, and MuSiQue 'was also used during method development' (Table 1 caption; Sec. 5.2, 5.5).
	D5	All questions come from public multi-hop datasets (MuSiQue, 2Wiki, HotpotQA) and a public 2026 enterprise benchmark, with no contamination check (Sec. 4.1, 5.5).
Xie et al. [256]	D5	Inputs are verbatim CLINC150, BANKING77 and TREC test texts with their public pre-2026 labels, which fix the required outputs, and no contamination check is reported (Sec. 3; Appendix A).
Xu et al. [260]	D5	The seven datasets (MMLU-Pro, SuperGPQA, MuSR, ToMBench, LAR-ECHR, SATA-Bench, BFCL V4) are public pre-2026 benchmarks used verbatim with gold answers and no contamination check is reported (Section 4.1; Appendix B.2).
Xu et al. [257]	D4	All rules receive the same inputs and release guards, but the relaxed signed-number baseline was added post hoc after inspecting verifier results and the verifier predicate was written after inspecting retained labels, so the evaluation is retrospective by the authors' account (Section III-A).
	D5	FinQA test and train splits are public 2021 releases with answer keys and the authors place training contamination outside the evaluation's scope (Section VI).
Yadav [261]	D3	The audited model is anonymized with no name, version or run date, and the 14 frontier models are named without versions or dates (Section 1 footnote; Section 4.2; Limitations).
	D4	HateBERT, Detoxify-unbiased, Perspective and the TF-IDF model were trained on Civil Comments labels that include the evaluation set, which the paper flags as contaminated (Sec. 3; Table 4).
	D5	Civil Comments is a public 2019 corpus with released rater fractions, the author states it may overlap training data, and the out-of-distribution set is not a contamination check (Limitations; Section 4.7).
Yamamoto and Kuroki [262]	D4	SummEval was used during method and probe development and is also the main evaluation, so the design was selected after seeing results on it; budgets are also unmatched (App.
	D5	All benchmarks are public pre-2026 sets (SummEval 2021 to JudgeBench 2025) with released labels, and the authors state contamination cannot be ruled out, with no check (Limitations).
Yan [263]	D3	Hosted Jev has no version or run date, and all baseline numbers are imported from the Chinese-Jev paper rather than rerun (Sec. 4.5).
	D5	CJ-Bench is built from public pre-2026 datasets, was evaluated in full for 14 checkpoints and six files, and the authors state no pretraining-contamination check was run (Sec. 7).
Yan [264]	D4	The aggregation rule and feature-subset size were selected after exploratory comparisons, and the authors say results need confirmation on an independent test partition (Sec. 4.3).

Study	Domain	Reason
	D5	AG News (2015) and TREC (2002) are public benchmarks with answer keys and no contamination check (Sec. 4.5).
Yang et al. [268]	D1	Worlds and protected mechanisms are AI-generated and screened by three LLM judges, with authored cost units, so the labels are synthetic (Sec. 4.1, 4.3; Appendix B.4).
	D3	Clef, Clef-flash and d1 are recorded only as clef, clef-flash and d1 with no version, and no run dates are given for most systems (Table 9; App.
Ye et al. [269]	D1	The headline reference is exact arithmetic on 64 author-generated expressions (seed 20260923) checked by an exact-rational evaluator, a synthetic procedure, with only three FRED index dates as external references (Section 4; Appendix B).
Yilmaz et al. [270]	D3	No version or date is given for Jev, whose numbers are quoted from Laya's report, and LAVOIR has only the moganai/lavoir weights with no revision or run date (Table 3 caption; Section 6).
You et al. [271]	D3	Jev is identified as jev-1.13.0 with access dates, but the compared DeepSeek model is served as deepseek-flash with no version identifier (Sec. 3.2; App.
	D5	HealthBench (2025) items are scored verbatim from the official open-source evaluation file, and the authors state training exposure is unknown with no contamination check (Sec. 3.5; Sec. 6).
Yu and Yao [272]	D5	GQA, SNLI-VE, TextVQA and TallyQA are public 2019 benchmarks and the backbone's pretraining data are undisclosed (note Test-set exposure).
Zhang et al. [277]	D5	Items are public pre-2026 benchmark records (BANKING77, 20 Newsgroups, MASSIVE) plus public JevBench and BEV subsets, and the paper reports no contamination analysis (Sec. 6, 7).
Zhang and Yu [281]	D5	All eight datasets are verbatim public benchmarks released before 2026 with answer keys and no contamination check, and HD has only 50 test items (Section 4.1; Table 1).
Zhang et al. [274]	D3	Hosted TypeSafe Jev has no version string or run date and CS1-4B has no revision, while the open comparators are named with reference implementations and only a few model-card revisions such as JEV-27B (Methods 0.13, references 24 and 25).
Zhang et al. [278]	D4	Components were chosen for JevCoder from component studies on samples drawn from the same 200+200 tasks that the full workflow is then scored on, so the system was selected after seeing results on its test tasks (Appendix D.4 Component selection for JevCoder).
Zhang [279]	D3	No version string, run date or endpoint identifier is given for Jev, although the VLM comparator has an OpenRouter ID (Section 2; Section 5.1).
	D5	AndroidWorld is a public benchmark whose tasks and checkers predate 2026 and no contamination check is reported (Section 5.1).
Zhang et al. [276]	D4	The text explicitly adds Sonnet, Haiku and Terra after inspection of earlier results (Sections 3.1 and 3.2, Limitations).
	D5	ContractNLI (2021) is public; 'an official held-out split does not establish absence from model training data' (Limitations).
Zhang et al. [275]	D5	All three benchmarks are public and predate 2026 and are used verbatim with their answer keys, with no contamination check reported (Section 2.2, Table 2; note Task and data).
Zheng and Chen [283]	D5	ProofWriter, LogiQA 2.0 and ARC-Challenge are public pre-2026 benchmarks used verbatim with their answer keys, evaluation items come from official training or test files, and no backbone contamination check is reported (Section 4.1; Appendix A).
Zhong et al. [285]	D1	Reference answers are computed by programmatic generators with exact ground truth, a synthetic procedure under the anchor, although free of annotation noise (Sec. 3.1, 4.1).

Study	Domain	Reason
	D3	No Jev version identifier or run date is given; the evaluated build is only implied by a documentation reference accessed 2026-09-25 (Sec. 3.2; reference list).
Zhou et al. [287]	D5	MMLU, GSM8K, AQuA, MultiArith, SVAMP, HumanEval and WebSRC are public benchmarks with released answers and no contamination check is reported (Section 5.1; note Exposure).
Zhu et al. [289]	D3	Jev is pinned as v1.13.0, but Table 3 lists Kev and Laya with no size, checkpoint or version although both come in several sizes or checkpoints (Table 3; refs. [17], [22]).
	D4	Intern-Decision is fine-tuned on a 17,880-sample corpus that includes the AG News, ToolACE, Typed Decision and WildJailbreak sources it is scored on, with temperatures fitted only for it, while baselines are used as is (Table 1; Table 4 caption; Appendix B).
	D5	Four of the seven suites are public pre-2026 datasets that are also training sources, public Jevbench 'has informed development', and the only overlap check covers the pilot (Sec. 4.2, 5.1).
Zhu et al. [288]	D1	The central DecisionGym benchmark is the authors' simulated desk whose answer key is derived from scripted plans at every reachable state, a simulator-computed synthetic reference, and the public BFCL and When2Call labels are not described (Section 3; Appendix D).
	D3	Jev is given as typesafe/jev-1.13, but the compared OpenAI hosted Decisions API is identified only as a service with no model or version (App.
Zhu et al. [290]	D4	Ablations run on the same complete test sets decided the adopted design, the MLP removal and the two-stage schedule, so settings were selected after seeing test results (Sec. 3.4; Tables 7 and 8).
	D5	Video-MME, LVBench, LongVideoBench, MLVU and the TimeLens-Bench grounding sets are public benchmarks predating 2026 and the training data come from the TimeLens lineage with no contamination analysis (Section 3.1; Section 2.4).
Zhuang et al. [292]	D3	Hosted Jev is listed only as Jev (hosted) with no version or date anywhere, the StartLux checkpoints carry no hash, Perplexity Decider v1.1 and Clef have access dates but no revisions, and vLLM is 0.31 in Table 5 but 0.19.1 in Appendix D (Table 4, Table 5, Appendix D.3).
	D4	The Decision Index public subset served as the development benchmark for choosing Stage 2 recipes, attention variants and seed averages, so the authors state its scores reflect model selection and are not independent test results (Section 4.1).
Hume, Archer [77]	D2	the 10,000-call headline does not reconcile with the itemized trial counts.
Palmer, Jared [163]	D3	Jev is unpinned ('live Jev' via /v1/systemone), with no version or date (note Decision models; Notes).
	D4	Kev is LoRA-trained plus a delta fine-tune on decision-v7 with a fitted temperature, compared against zero-shot Jev on development partitions (note Comparators).
LocalLLaMA [132]	D1	Soft gold is the mean of three samples from a roughly 4B teacher at temperature 0.7; the card says it 'does not measure correctness' (How it was built; Limitations).
Balian, Rob [9]	D4	Rev is LoRA-trained on the train splits of the same five datasets against zero-shot Jev; JevBench misses were read to design synthetic data (README; Limitations).
TypeSafe AI [230]	D1	Reference labels are the average of GPT-6 Astra and Claude Fable 5.1 answers ('Assume the harness is correct'), a model consensus (page).
	D2	No item counts, sample sizes or intervals for any point (note n).
	D3	Jev version not printed and no dates (note Decision models).

Study	Domain	Reason
Rauch [185]	D2	No n, task definition or protocol; '~5-18x' is a range quoted at its ceiling (note Cost and latency).
	D3	No version string for Jev (note Decision model).

S4 Decision-model card template and full reporting checklist

Table S4 gives every field of the decision-model card of Section 11, with an example quoted from the census card of Laya, an open decision model under an Apache license, as of its access date [39]. A card filled this way lets a caller rebuild the decision contract record without the original repository. The multilingual Laya checkpoint uses an mmBERT-base encoder rather than the ModernBERT-large encoder of the English one and has a different parameter count, so a card names one checkpoint and revision, not a family [38].

Table S4. The decision-model card template, with the content each field holds and an example drawn from the Laya card.

Field	Content	Example
Name.	The model's given name.	"Laya."
Vendor or author.	The organization or person publishing the checkpoint.	"Convai Innovations."
Release date.	The date the checkpoint was published.	"2026-09-18."
Access date.	The date the card was read for this survey.	"2026-09-24."
Availability and license.	Open or hosted, and the license terms.	"Open weights, Apache-2.0."
Backbone.	The pretrained encoder or decoder and its width.	"ModernBERT-large, 395M, bidirectional, fully fine-tuned."
Parameter count.	The total parameter count, as reported by the weights file.	"421M total (safetensors.total 421,293,830)."
Decision head.	The layers and readout that turn the backbone's output into a distribution.	"2 transformer layers, an option-marker scorer, and an act/escalate head, trained from scratch; every option scored at its own [MASK] token, softmaxed over that question's options."
Training objective.	The loss or reward the checkpoint was trained against.	"RLCD: policy reports a distribution, reward is a strictly proper scoring rule (log + spherical, plus ranked probability score for ordinal questions), REINFORCE with group-mean baseline (GRPO-style)."
Training data.	The corpora or generation procedure behind training.	"Not itemized by source/size in the card beyond the RLCD training description and the separate typed-decisions fine-tuning split."
Primitives.	The declared answer-space types the model accepts.	"choice, score, noul."
State modality and context length.	What kind of input the state may be, and the token budget.	"text, email, ticket, or JSON. 512 tokens (head_max_len = 192)."
Cardinality limit.	Any disclosed cap on the number of options, or the observed degradation without one.	"Not disclosed as a fixed cap; accuracy degrades sharply above ~20 options at default settings (Banking77, 77 options, scores 0.425)."

Field	Content	Example
Latency with hardware.	A timing value and the hardware it was measured on.	"39.5 ms p50 for 1 question on a Tesla T4 GPU (English checkpoint)."
Price.	The cost per call or per token, or its absence for a self-hosted model.	"$0 self-hosted (open weights)."
Confidence definition.	The formula behind any confidence value the model returns.	"Not disclosed as a formula in this card beyond the reported probabilities themselves."
Probability grid.	The precision or rounding of the returned probabilities.	"Not disclosed."
Calibration as shipped, with source.	A calibration value and the evaluation that produced it.	"Raw ECE 0.466; refitting one temperature per (question type, option count) moves mean ECE to 0.081."
Disclosed failure modes.	Weaknesses the publisher itself states.	"Base checkpoint near chance zero-shot on typed-decisions (0.362); Khmer collapses to 0.000 accuracy at 0.952 confidence; noul can follow its label tokens instead of the state; action.act_probability carries no usable signal (AUROC 0.30)..."
Evaluation protocol published.	Whether a benchmark script or methodology accompanies the card.	"Yes; BENCHMARKS.md on GitHub and a typed-decisions in-card table (400 cases, 2,000 decisions) with Jev figures marked as third-party published, not measured by Laya's authors."
Source URLs.	The card and repository locations the entries above were drawn from.	https://huggingface.co/convaiinnovations/laya/raw/55cf4c4ebb4ebe31b2550e8bdf3bd21b99753851/README.md and https://github.com/NandhaKishorM/laya.

Table S5 gives every item of the standard in its five groups and, where one applies, the ledger field or the risk-of-bias domain of Wolff et al. [250] and Whiting et al. [249] that it operationalizes. Ledger columns follow the schema of Section 7, which records for each extracted number the study, model and version pin, reference-label type, comparator and tuning budget, calibration quantity and cascade result.

Table S5. The full reporting checklist for decision-model evaluations.

Item	What to report	Explanation
What to pin		
Model and revision	Provider or checkpoint, resolved revision and the date it was confirmed.	A name can serve another checkpoint later, hence the version_pin column and the version-pinning domain.
Serving temperature	Decoding temperature and any other sampling parameter active at evaluation.	A release-page value need not be the service default, so unstated temperatures can put two predictors under one name.
Schema and identifiers	Literal question text, option identifiers and the rubric of each, as separate fields.	Proposition 5.5 splits an option's logit into rubric and name terms, and only separate fields allow the name-swap test.
Probability grid	Smallest positive returnable value and share of returned zeros.	Proposition 5.2 ties this pair to a floor on calibration error for rare events, before model quality enters.
State encoding	State serialization, token budget and any truncation rule.	A tight per-option budget can cost accuracy for reasons unrelated to decision quality.
What to sample		

Continued on the next page

Table S5, continued

Item	What to report	Explanation
Reference-label provenance	Human adjudicated, single human rating, crowd rating, teacher model, model consensus or synthetic procedure.	Hence the reference_label_type column and the label-independence domain, since a teacher or synthetic label is not independent evidence for a model trained toward that source.
Sample size and exposure	Item count, grouping against near-duplicate leakage and whether the set predates the cutoff.	Public pre-cutoff items may sit in training data, and an ungrouped split can inflate agreement.
Calibration quantity per primitive		
Choice calibration	Classwise calibration or reliability diagram beside any top-label accuracy.	Top-label calibration can hold while classwise calibration fails on rarer classes, which a top-label value never examines.
Score calibration	Ranked probability score or cumulative calibration at each ordinal threshold, beside any expected-level value.	Proposition 5.6 gives two forecasts with one expected level that disagree on every cumulative probability.
Joint-question calibration	Fréchet and Hoeffding interval or a stated dependence model for a conjunction of isolated answers, not the product of marginals.	That product asserts an independence the contract never supplied (Proposition 5.1).
Confidence statistic	Confidence formula, the answer-set size used and whether sizes were compared.	Proposition 5.3 shows the accepted-answer error bound grows with answer-set size.
Cascade and review results		
Cascade recall	Screening forwarding rate and downstream detection rate, as two values rather than one accuracy.	Proposition 5.4 gives the final recall as their product. The cascade_result field records both, so forwarded-only scores are not read as whole-pipeline recall.
Review band and cost	Routing threshold or band, assumed costs of a false positive, a false negative and a review, and resulting review volume.	The band of Proposition 5.4 is non-empty only under a stated cost relation.
Speed and cost claims	Latency and cost with hardware, batch size and concurrency, each marked as vendor claim or independent measurement.	A ratio shifts with workload, batch size and provider.
Handling version churn		
Revision-change protocol	A rule for a mid-study change of served checkpoint or default parameter, including which rows are re-run.	Otherwise rows from before and after the change can be pooled under one name.
Superseded values	Both values when a revision change causes disagreement, each with its revision, date and settlement.	This keeps the audit trail that one overwritten number would remove.

S5 Full census of original implementations

Table S7 lists every entry that Section 6 counts as an original implementation, with its parameter count, base model and decision head, and Table S6 condenses this list to the fifteen families with an independent card. Table S8 breaks the derivative count down by parent checkpoint. Every value

is read from the census files of the companion repository at release v1.2.2, which also record each entry's creation date and family at the cutoffs stated in Section 2.

Table S6. The ecosystem census, condensed to the notable families with an independent card (Section 6), with every cell abbreviated from the family's own card or repository.

Family	Author	First release	Backbone and size	Head	Objective	Primitives	Temperature or calibration as shipped
Jev 1.13 [224]	TypeSafe AI	15 Sep 2026	Undisclosed; hosted only	Undisclosed; a shared prefix with isolated per-question branches is inferred [77] (inference)	RLCD, an unpublished reinforcement-learning recipe [221]	Choice, Noul, Score	Confidence statistic not documented as a formula; a vendor demo approximates $(Kp_{\max} - 1)/(K - 1)$ [223]
Laya (English, multilingual) [38, 39]	Convai Innovations	18–19 Sep 2026	ModernBERT-large 395M (English); mmBERT-base 307M (multilingual)	Two-layer option-marker scorer plus an act/escalate head, trained from scratch	RLCD: a proper scoring rule (log, spherical, ranked probability score) with a GRPO-style policy update	Choice, Score, Noul	No single shipped default; one temperature refit per question type and option count moves raw ECE 0.466 to 0.081 (English)
Kev [161–163]	Jared Palmer	18–19 Sep 2026	Qwen2.5, Qwen3 and Qwen3.5 backbones, 0.5B to 9B	LoRA adapter (rank 16) plus a pointer head reading option logits, no free-text generation	Cross-entropy on the option distribution, the decision-v6/v7 family recipe	Choice, Noul, Score	T = 2.35 (0.8B) and T = 2.30 (9B), each fitted on its own checkpoint; the other sizes are not independently re-verified
Bespoke-Nimble-9B [13, 14]	Bespoke Labs	18 Sep 2026 (checkpoint original-2676)	Qwen3.5-9B	LoRA adapter (rank 16) scoring the allowed answer tokens directly	Cross-entropy over allowed answer-token logits	Choice, Score, Noul	T = 2.179 as fitted; ECE on a held-out set moves 0.052 to 0.054 as served (developer README)
this-that-model [59], a fine-tune of decider-2b	FLock.io	20–23 Sep 2026 (1.0 to 1.2)	decider-2b backbone (Qwen3.5-style hybrid, 1.9B)	Inherits decider-2b's letter-logit readout; no independently trained head is disclosed	Not disclosed beyond the decider-2b lineage	Choice-style typed decisions	Not disclosed for 1.0 to 1.1; 1.2 reports 0.985 accuracy and 0.009 Brier on a frozen third-party cohort
decider [147]	Mapika	16–22 Sep 2026	Qwen3.5 backbones, 0.8B to 35B-A3B	Letter-logit readout from the LM head, divided by a stored temperature, no separate trained head	Supervised cross-entropy stages v1–v8; PPO with a proper log score and a KL constraint, v8–v10 (2B only)	Noul, Choice, Score	T = 1.30 (2b, v10); the other sizes report stored temperatures from 1.03 to 1.94

Family	Author	First release	Backbone and size	Head	Objective	Primitives	Temperature or calibration as shipped
OpenThai-SystemOne [79]	iApp / OpenThaiGPT	20 Sep 2026	Qwen3.5-0.8B, continued-pretrained on Thai text	256-way slot head replacing the LM head, read at one control token	Supervised decision conversion of public classification and QA data, plus a Brier-loss calibration stage	Choice (up to 255), Score, Noul	Per-type learned temperatures: choice 1.055, noul 1.047, score 1.008
Von [248]	wfzyx (independent)	19 Sep 2026	ModernBERT-large 395M	Option-marker head with an order-invariant attention mask (v1.2)	Retrained from the v1.1 weights under the new attention mask	Choice-style decisions; Noul with a fitted prior correction	Input-conditioned calibration map rather than one scalar temperature; ECE 0.045 to 0.109 across difficulty tiers
Decision-1.0-Sol-2B [131]	llm-semantic-router	21 Sep 2026	Qwen3.5-2B	Shared candidate head reading candidate endpoints and the final query vector	Not itemized by loss or recipe in the card	Choice (2 to 255), Noul, Score	Not disclosed; the card reports 66.32% weighted accuracy only
Open-Jev [20]	Zefan Cai	20 Sep 2026	Qwen3.5-9B	LoRA adapter (rank 8) plus a scalar head initialized from a Yes/No readout	20,204 optimizer steps over one pass of a frozen release-v2 training split	Choice, Noul, Score	T = 1.897, fitted on 512 calibration rows; test-split ECE 0.0077, out-of-domain ECE 0.0374
open-jev-deberta [90]	kotoba-lang	18 Sep 2026	DeBERTa-v3-large	Three-layer scoring head over pooled question and option representations	Cross-entropy plus Brier loss, with gold-preserving question augmentation	Choice (up to 255), Score, Noul	Post-hoc temperature fitted on a validation split, value not restated in the card; in-domain ECE 0.022
pngwn scorer [172]	pngwn (independent)	16 Sep 2026	Qwen3.5-4B	LoRA (rank 16) plus a new scalar score head; sequence-classification readout per option	Not itemized beyond the scoring-head and LoRA description	Noul, Choice, Score	T = 1.75, fitted on a validation split; ECE 0.135 to 0.044 as served
Verdict [71]	Heman10x-NGU	17 Sep 2026	ModernBERT plus GLiClass, 151M	Non-autoregressive typed-question scorer, GitHub-only	RLCD with a Brier-loss term, per the repository description	Choice, Score, Noul	Not disclosed as a scalar; the repository reports calibrated confidence values

Family	Author	First release	Backbone and size	Head	Objective	Primitives	Temperature or calibration as shipped
SemIf [219]	TheoLeeCJ	15 Sep 2026 (as Open-Jev)	MiniCPM5-2B, Qwen3.5-4B, Qwen3.8-27B	Reads option probabilities directly from a frozen backbone, GitHub-only	No generation; per-workload temperature calibration added after the rename	Choice, Noul	Per-workload temperature calibration; value not restated in the repository card
OpenDecision [232]	deepanwadhwa	17 Sep 2026	ModernBERT-large zero-shot v2.0	Zero-shot classification head served through a FastAPI, vendor-compatible endpoint	Not itemized; a packaged zero-shot backend rather than a newly trained recipe	Choice, Noul, Score, Relation	Not disclosed; the repository documents the API contract rather than a calibration fit
Jev 1.14 [222, 227]	TypeSafe AI	14 Oct 2026	Undisclosed; hosted only, same price and context as 1.13	Undisclosed	RLCD, as for 1.13 (vendor wording unchanged)	Choice, Noul, Score, Box	Confidence formula now published for each primitive; Noul returns no confidence
Microsoft-Decision-1 [151, 152]	Microsoft	9 Oct 2026	Post-trained Qwen3.5-9B; hosted in Foundry and on OpenRouter	Undisclosed; single-pass scoring of every supplied option	Not disclosed	Choice, Noul, Score	Calibration goal stated; documentation advises reading Score as an ordering, not a calibrated rating
Trio-Spark [145]	MachineFi	30 Sep 2026	Undisclosed; hosted only	Undisclosed; one pass returns the chosen option and a probability for each	Not disclosed	Choice, 2 to 8 options	Vendor ECE 0.019 (10 bins) on its own panel, not claimed to transfer
Clef [37]	Cloudflare	30 Sep 2026	Qwen3.8-27B, open weights (Apache) and hosted	Joint schema head, a small transformer over the final hidden states scoring all options of all questions	Label-smoothed cross-entropy with a Brier loss term	Choice, Score, Noul	Not disclosed; per-question softmax with no stated temperature
pplx-decider [171]	Perplexity	1 Oct 2026	Qwen3.8-27B, open weights (Apache)	A [255, 5120] decision head over the backbone, full-attention layers run noncausally	Not disclosed; training data converted into Choice and Noul decisions	Choice, Noul	One saved calibration temperature, value not stated

Table S7. Every T1 original implementation in the census (Section 6), in order of creation, with its parameter count, base model and decision-head type as in Figure 3. A dagger marks entries added by the second sweep, a dash an unreported value, and u an undisclosed head. Creation dates and families are in the companion repository.

Repository	Parameters	Base model	Head
llm-semantic-router/Decision-1.0-Eos-0.8B	–	–	b
llm-semantic-router/Decision-1.0-Kai-0.6B	–	–	b
llm-semantic-router/Decision-1.0-Lex-0.6B	–	–	b
Mapika/decider-2b	1.88B	Qwen/Qwen3.5-2B-Base	c
pngwn/system-one-qwen3.5-4b-scorer	–	Qwen/Qwen3.5-4B-Base	d
pngwn/nanodiff-350m-typed-decisions-lam1	–	Sebasdi/nanodiff-350m-base	c
DavidHatley/system-one-mini	69.3M	distilbert/distilbert-base-uncased	a
Mapika/decider-2b-vision	2.21B	Qwen/Qwen3.5-2B-Base	c
mobarmg/jev-schema-scorer-deberta-v3-large	435M	microsoft/deberta-v3-large	d
bespokelabs/Bespoke-Nimble-9B	–	Qwen/Qwen3.5-9B	c
SargeDev/jev-distill-corpus	–	–	c
convaiinnovations/laya	421M	–	b
com-kotobalabs/open-jev-deberta-v3-large	434M	microsoft/deberta-v3-large	b
Mannedood/local-system-one-student	149M	answerdotai/ModernBERT-base	a
jaredpalmer/kev-0.5b	–	Qwen/Qwen2.5-0.5B	b
larkooo/gemma-e2b-rlcd†	–	google/gemma-4-E2B-it	c
jaredpalmer/kev-0.6b	–	Qwen/Qwen3-0.6B-Base	b
convaiinnovations/laya-multilingual	322M	–	b
jaredpalmer/kev-4b	–	Qwen/Qwen3.5-4B-Base	b
jaredpalmer/kev-8b	–	Qwen/Qwen3-8B-Base	b
anthonym21/qwen3-0.6b-rlcd-decision†	–	Qwen/Qwen3-0.6B-Base	c
dwidlee/systemone-lite-0.5b	494M	Qwen/Qwen2.5-0.5B-Instruct	c
Mapika/decider-0.8b	752M	Qwen/Qwen3.5-0.8B-Base	c
AndeyTait/JevForge-0.8B	–	Qwen/Qwen3.5-0.8B	d
wfzyx/von-1.0	395M	answerdotai/ModernBERT-large	b
argos1111/modernbert-ja-310m-jev	315M	sbintuitions/modernbert-ja-310m	d
lafalce/system-one-model	–	answerdotai/ModernBERT-base	b
lewislululu/jevon	–	–	b
shreyanbr/system-one-zeroshot	70.8M	MoritzLaurer/deberta-v3-xsmall-zeroshot-v1.1-all-33	d
shreyanbr/system-one-distilled	70.8M	MoritzLaurer/deberta-v3-xsmall-zeroshot-v1.1-all-33	d
shreyanbr/system-one-gold	70.8M	MoritzLaurer/deberta-v3-xsmall-zeroshot-v1.1-all-33	d
vagmi/jev-lite	–	google/gemma-4-E4B-it	c
Mapika/decider-35b-a3b	34.7B	Qwen/Qwen3.5-35B-A3B-Base	c
Nebulaw1/jev-qwen3.5-4b-legal-lora	–	Qwen/Qwen3.5-4B-Base	c
Nebulaw1/jev-qwen3.5-0.8b-legal-lora	–	Qwen/Qwen3.5-0.8B-Base	c
iapp/OpenThai-SystemOne	753M	Qwen/Qwen3.5-0.8B-Base	a
AstroHan/decision-head-qwen3.5-4b-rlcd-32k†	–	Qwen/Qwen3.5-4B	c
altslate/certo-decision-model	–	–	b
azharmo/build-jev-from-scratch	–	–	a
ZefanCai/Open-Jev-2B	–	Qwen/Qwen3.5-2B	c
ZefanCai/Open-Jev-9B	–	Qwen/Qwen3.5-9B	c

Repository	Parameters	Base model	Head
akhilaaa3/Jev-Omni	–	google/gemma-4-12B-it	a
jaredpalmer/kev-9b	–	Qwen/Qwen3.5-9B-Base	b
jaredpalmer/kev-0.8b	–	Qwen/Qwen3.5-0.8B-Base	b
samatv256/mini-Jev	263K	Qwen/Qwen3-0.6B	b
Praveenrajus/jevify-qwen3.5-4b	–	Qwen/Qwen3.5-4B	b
Praveenrajus/jevify-qwen3.5-2b	–	Qwen/Qwen3.5-2B	b
SargeDev/jev-gate-student-b	–	–	c
SeanLiu/Jev-Vision	–	Qwen/Qwen3-VL-8B-Instruct	b
llm-semantic-router/Decision-1.0-Nox-4B	–	Qwen/Qwen3.5-4B	b
llm-semantic-router/Decision-1.0-Sol-2B	–	Qwen/Qwen3.5-2B	b
Pdbz199/local-decision-model	–	answerdotai/ModernBERT-base	b
aimeigaoshou/agent-jev	598M	–	b
abidlabs/jev-typed-decisions-causal-0.6b	–	Qwen/Qwen3-0.6B-Base	c
mogita/jev-decider-qwen3-4b	–	Qwen/Qwen3-4B	c
lostargon/Tiny-Jev	596M	Qwen/Qwen3-0.6B	b
marcmagn1/kev-08b-typed-v1	–	–	b
Praveenrajus/jevify-qwen3-vl-2b	–	Qwen/Qwen3-VL-2B-Instruct	c
kaivoss/system-one-270m	268M	unsloth/gemma-3-270m-it	c
Praveenrajus/jevify-qwen3.5-4b-t2	–	Qwen/Qwen3.5-4B	b
Praveenrajus/jevify-qwen3-vl-2b-t2	–	Qwen/Qwen3-VL-2B-Instruct	c
top7777/jev-schema-scorer-deberta-v3-large	435M	microsoft/deberta-v3-large	d
llm-semantic-router/Decision-1.0-Lux-9B	–	Qwen/Qwen3.5-9B	b
sshalimov04/open-jev-base	–	jhu-clsp/mmBERT-small	d
jaswanthsanjay88/rev-decision-model	421M	–	b
olafura/gemma4-12b-system-one	–	google/gemma-4-12B-it-qat-w4a16-ct	u
IJyad/jeb-typed-decisions	178M	IJyad/jeb	b
mghafiri/qwen3.5-0.8B-decision-model	752M	Qwen/Qwen3.5-0.8B-Base	c
shreyanbr/nev-lite-systemone	–	sentence-transformers/all-MiniLM-L6-v2	d
alibiserikbay/JevK5	4.21B	Qwen/Qwen3.5-4B	c
Mapika/decider-4b	4.21B	Qwen/Qwen3.5-4B-Base	c
ait-hf/certus-jev-like-v0007	–	Qwen/Qwen2.5-1.5B-Instruct	c
tasksource/modernbert-tasksource-jev	153M	–	b
tianxinwei/JevAny-27B-SFT	–	Qwen/Qwen3.8-27B	b
tianxinwei/JevAny-27B-RLCR	–	Qwen/Qwen3.8-27B	b
Praveenrajus/jevify-qwen3.5-4b-t2-lowlr	–	Qwen/Qwen3.5-4B	b
FluidInference/kev-0-5b-coreml	–	–	b
vigneshlabs/ballot-jev-0.5b	–	Qwen/Qwen2.5-0.5B	b
ZefanCai/Open-Jev-27B-v1.1	–	Qwen/Qwen3.8-27B	c
asjson/jevson-4b-01	–	Qwen/Qwen3-4B	u
SUPER321/jevflash-doom-basic-0.6b	–	Qwen/Qwen3-0.6B-Base	b
0xSojalSec/Jev-Omni	–	google/gemma-4-12B-it	a
bespokelabs/Bespoke-Nimble-9B-v2†	–	Qwen/Qwen3.5-9B	c
guanxuyu/visual-jev-4b-answer-sft	–	Qwen/Qwen3-VL-4B-Instruct	c
DawoodKMasood/jev-gemma-3-270m	–	google/gemma-3-270m-it	c
autotrust/JEV	7.94B	Qwen/Qwen3.5-9B	a
autotrust/JEV-9B†	–	Qwen/Qwen3.5-9B	a
xuhaodev/Qwen3-1.7B-Jev	–	Qwen/Qwen3-1.7B	c
soyrsoyr/jev-playground-rlcd	–	MoritzLaurer/deberta-v3-large-zeroshot-v2.0	d

Repository	Parameters	Base model	Head
TokenRhythm/NeoHorse-Jev-4B	–	TokenRhythm/NeoHorse-1-4B	b
chaoliangUNSW/Jev-Style-Qwen3.5-2B-Decision-v2	1.88B	Qwen/Qwen3.5-2B-Base	c
JohnP1/kev-gemma4-e2b	–	google/gemma-4-E2B	b
DKNTZMN/gan8-vs-jev	–	–	a
wwydmanski/bielik-minitron-jev-v0.1	–	speakleash/Bielik-Minitron-7B-v3.0-Instruct	b
lostargon/Tiny-Jev-1.7B	1.72B	Qwen/Qwen3-1.7B	b
wwydmanski/bielik-minitron-jev-v0.2	–	speakleash/Bielik-Minitron-7B-v3.0-Instruct	b
alibiserikbay/JevK5-2B	1.88B	Qwen/Qwen3.5-2B	c
SargeDev/Jev_Qwen3.8-27B	26.9B	huihui-ai/Huihui-Qwen3.8-27B-abliterated	u
guoan1/jev-qwen35-08b	–	Qwen/Qwen3.5-0.8B-Base	c
Fr0zencr4nE/jev-spatial	–	allenai/Molmo2-ER	a
jaredpalmer/kev-27b†	–	Qwen/Qwen3.8-27B	b
shgao/rsi-jev-v1.0-qwen3.5-0.8b†	–	Qwen/Qwen3.5-0.8B-Base	b
Rydward98/Jev-Omni	–	google/gemma-4-12B-it	a
wwydmanski/bielik-minitron-jev-v0.3	–	speakleash/Bielik-Minitron-7B-v3.0-Instruct	b
VTXAI/VTX-JEV-1	7.13M	–	b
divyanshx11/JEVision	–	Qwen/Qwen3.5-0.8B-Base	b
joyfox/Qwen3.5-0.8B-JEV	–	Qwen/Qwen3.5-0.8B	b
Quazim0t0/Byrne-Jev-79M	–	–	b
chaoliangUNSW/Jev-Style-0.8B-Decision-v3†	–	Qwen/Qwen3.5-0.8B	c
Rckflr/kev-gemma-270m-experimental†	–	google/gemma-3-270m	b
flymy-ai/decision-4b-v1.1†	–	Qwen/Qwen3.5-4B	c
flymy-ai/decision-4b-v1.2†	–	Qwen/Qwen3.5-4B	c
Praveenrajus/jevify-qwen3.5-4b-readout-coh†	–	Qwen/Qwen3.5-4B	c
Praveenrajus/jevify-qwen3.5-4b-readout†	–	Qwen/Qwen3.5-4B	c
Praveenrajus/jevify-qwen3.5-4b-base-readout-coh†	–	Qwen/Qwen3.5-4B-Base	c
Praveenrajus/jevify-gemma-4-e4b-it-readout-coh†	–	google/gemma-4-E4B-it	c
Praveenrajus/jevify-qwen3.5-2b-readout†	–	Qwen/Qwen3.5-2B	c
Praveenrajus/jevify-qwen3.5-2b-readout-coh†	–	Qwen/Qwen3.5-2B	c
Praveenrajus/jevify-qwen3.5-2b-base-readout†	–	Qwen/Qwen3.5-2B-Base	c
Praveenrajus/jevify-qwen3.5-4b-base-readout†	–	Qwen/Qwen3.5-4B-Base	c
Praveenrajus/jevify-qwen3.5-9b-readout†	–	Qwen/Qwen3.5-9B	c
Praveenrajus/jevify-gemma-4-e4b-it-readout†	–	google/gemma-4-E4B-it	c
Praveenrajus/jevify-qwen3.5-2b-base-readout-coh†	–	Qwen/Qwen3.5-2B-Base	c
sunziqin/millennium-jev-0.5b†	–	Qwen/Qwen2.5-0.5B-Instruct	b
alibiserikbay/JevK5-9B†	–	Qwen/Qwen3.5-9B	c
Praveenrajus/jevify-gemma-4-e4b-readout†	–	google/gemma-4-E4B	c
sdmlai/nano-jev†	–	microsoft/MiniLM-L12-H384-uncased	d
skundu42/kev†	–	jhu-clsp/ettin-encoder-400m	d
Praveenrajus/jevify-qwen3.5-9b-readout-coh†	–	Qwen/Qwen3.5-9B	c
sunziqin/millennium-jev-1.5b†	–	Qwen/Qwen2.5-1.5B-Instruct	b
Praveenrajus/Llama-3.1-Tulu-3-8B-DPO-jevify-readout†	–	allenai/Llama-3.1-Tulu-3-8B-DPO	c

Repository	Parameters	Base model	Head
Praveenrajus/Llama-3.1-Tulu-3-8B-DPO-jevify-readout-coh†	–	allenai/Llama-3.1-Tulu-3-8B-DPO	c
Praveenrajus/Llama-3.1-Tulu-3-8B-SFT-jevify-readout†	–	allenai/Llama-3.1-Tulu-3-8B-SFT	c
HIT-TMG/JevEmbed-KaLM-Embedding-V2.5†	–	KaLM-Embedding/KaLM-embedding-multilingual-mini-instruct-v2.5	d
Praveenrajus/Llama-3.1-Tulu-3-8B-SFT-jevify-readout-coh†	–	allenai/Llama-3.1-Tulu-3-8B-SFT	c
Praveenrajus/jevify-smollm3-3b-sft-readout†	–	HuggingFaceTB/SmolLM3-3B-checkpoints	c
Praveenrajus/jevify-smollm3-3b-sft-readout-coh†	–	HuggingFaceTB/SmolLM3-3B-checkpoints	c
Praveenrajus/jevify-smollm3-3b-apo-readout†	–	HuggingFaceTB/SmolLM3-3B-checkpoints	c
Praveenrajus/jevify-smollm3-3b-apo-readout-coh†	–	HuggingFaceTB/SmolLM3-3B-checkpoints	c
ryandcunha/mmbu-jev-qwen3-0.6b-closed†	–	Qwen/Qwen3-0.6B	b
nouton/jev-ru-4b-lora-v2†	–	Qwen/Qwen3.5-4B	c
NAME0x0/Jevlet†	–	BAAI/bge-small-en-v1.5	b
sidman007/jev-vlm-4b-lite-gguf†	–	Qwen/Qwen3.5-4B-Base	c
HIT-TMG/JevEmbed-Qwen3-Embedding-0.6B†	–	Qwen/Qwen3-Embedding-0.6B	d
shgao/rsi-jev-v2.0-qwen3.5-2b†	–	Qwen/Qwen3.5-2B-Base	b
per2021/kev-2b-qwen3.5†	–	Qwen/Qwen3.5-2B-Base	b
aravind1706/decider-ettin†	–	jhu-clsp/ettin-encoder-68m	b
jaluus/jevpertus-v1.5-8B†	–	swiss-ai/Apertus-v1.5-8B	b
jaluus/jevpertus-8B-Instruct-2509†	–	swiss-ai/Apertus-8B-Instruct-2509	b
hiteshluke/zyot-decider-12b-adapter†	–	–	u
chaoliangUNSW/Jev-Style-2B-Decision-v3†	–	Qwen/Qwen3.5-2B	c
dr3x1/jevsec†	–	XHToken/Spark-X2.5-4B	u
shgao/rsi-jev-v2.1-qwen3.5-2b†	–	Qwen/Qwen3.5-2B-Base	b
Quazim0t0/On-Fly-Jev†	–	–	b
abhishek085/jev-control-core†	–	Qwen/Qwen3.5-0.8B-Base	c
abhishek085/jev-control-es†	–	answerdotai/ModernBERT-base	b
ytbai/MiniCPM5-2B-Jev†	–	openbmb/MiniCPM5-2B	a
rongxinzy/LightJev-0.6B-typed-decisions†	–	Qwen/Qwen3-0.6B	b
tasksource/tasksource-decider-nano†	–	cross-encoder/ettin-reranker-150m-v1	b
sainath/tiny-jev-support-v03-supervised†	–	Qwen/Qwen3-1.7B	d
avbiswas/bev-decider-0.4B†	–	Qwen/Qwen3-0.6B	b
shgao/rsi-jev-v3.0-qwen3.5-2b†	–	Qwen/Qwen3.5-2B-Base	b
BricksDisplay/jevling-0.8b-v0.1†	–	Qwen/Qwen3.5-0.8B	c
BricksDisplay/jevling-e2b-v0.1†	–	google/gemma-4-E2B-it	c
MSGEncrypted/ocf-typed-decisions-mbert-base†	–	answerdotai/ModernBERT-base	b
flymy-ai/decision-4b-v1.4†	–	Qwen/Qwen3.5-4B	c
vincentzhou/jev-student-4b-yizao†	–	Qwen/Qwen3.5-4B-Base	c
hotchpotch/bekko-system-one-v0-17m†	–	cross-encoder/ettin-reranker-17m-v1	b
mohdyaser/jev-multimodal†	–	mlx-community/gemma-4-e4b-it-4bit	c
startlux-models/StartLux-Decision-v1-0.8B†	–	–	c
startlux-models/StartLux-Decision-v1-2B†	–	–	c
startlux-models/StartLux-Decision-v1-4B†	–	–	c
startlux-models/StartLux-Decision-v1-27B†	–	–	c
startlux-models/StartLux-Decision-v1-9B†	–	–	c

Repository	Parameters	Base model	Head
autotrust/JEV-Gemma4-26B-A4B†	–	google/gemma-4-26B-A4B-it	a
Mapika/decider-12b†	–	google/gemma-4-12B-it	c
Mapika/decider-chat-gemma4-31b†	–	google/gemma-4-31B-it	c
Mapika/decider-chat-qwen3.6-27b†	–	Qwen/Qwen3.6-27B	c
flymy-ai/decision-4b-v1.5†	–	Qwen/Qwen3.5-4B	c
CVLTAI/s1-jev-cvltist†	–	Qwen/Qwen3.5-4B	c
seongyeon1/dlm-jev-tickets-lora†	–	mlx-community/diffusiongemma-26B-A4B-it-4bit	c
bespokelabs/Bespoke-Nimble-9B-v3†	–	Qwen/Qwen3.5-9B	c
autotrust/JEV-27B-VL†	–	Qwen/Qwen3.8-27B	a
jevhome/jevhome-E†	–	ibm-granite/granite-embedding-small-english-r2	d
jevhome/jevhome-B†	–	answerdotai/ModernBERT-base	b
jevhome/jevhome-L†	–	answerdotai/ModernBERT-Large-Instruct	b
jevhome/jevhome-ettin-1b†	–	jhu-clsp/ettin-encoder-1b	b
Maincode/Matilda-Jev-v1†	–	Qwen/Qwen3.8-27B	a
TerryGao9/zh-decision-mmbert-v1b†	–	jhu-clsp/mmBERT-base	b
BricksDisplay/jevling-0.8b-v1†	–	Qwen/Qwen3.5-0.8B	c
StrandsAgents/strands-decider-2B-hobson-v19†	–	Qwen/Qwen3.5-2B-Base	b
Cloudflare/clef†	–	Qwen/Qwen3.8-27B	b
Cloudflare/clef-flash†	–	Qwen/Qwen3.5-9B	b
shgao/rsi-jev-v4.0-vl-qwen3.5-2b†	–	Qwen/Qwen3.5-2B-Base	b
startlux-models/StartLux-Decision-v1-35B-A3B†	–	–	c
sennaLLMLearner/qwen3-4b-system-one-lora†	–	Qwen/Qwen3-4B-Instruct-2507	c
muhammadnoman76/jev-urdu†	–	jhu-clsp/mmBERT-base	b
hiteshluke/zyot-decider-v3-verbalizer-4b†	–	unsloth/gemma-3-4b-it	c
perplexity-ai/pplx-decider-v1-27b†	–	Qwen/Qwen3.8-27B	a
xuhaodev/Qwen3.5-4B-Jev†	–	Qwen/Qwen3.5-4B	c
smha1012/jev-27b†	–	Qwen/Qwen3.8-27B	a
Praveenrajus/jevify-qwen-agentworld-35b-a3b-readout-coh†	–	Qwen/Qwen-AgentWorld-35B-A3B	c
carrtesy/EXAONE-4.0-1.2B-JEV†	–	LGAI-EXAONE/EXAONE-4.0-1.2B	c
torchcast-ai/torchcast-decision-12b†	–	google/gemma-4-12B-it	c
nickprock/archai-jev-qwen-1.5b†	–	Qwen/Qwen2.5-1.5B-Instruct	c
projectj/Flight-Autopilot-Decision-Model-Maverick1.0†	–	Qwen/Qwen3-0.6B	c
Kestrelyn/kestrel-decider-230m†	–	LiquidAI/LFM2.5-230M-Base	b
shgao/rsi-jev-v5.0-vl-3b†	–	Qwen/Qwen3.5-4B-Base	b
nickprock/archai-jev-zagreus-0.4b-ita†	–	mii-llm/zagreus-0.4B-ita	c
systemonedev/kenning-large-v0.4†	–	MoritzLaurer/deberta-v3-large-zeroshot-v2.0-c	d
BricksDisplay/jevling-e2b-v1†	–	google/gemma-4-E2B-it	c
AmirBraham/jevstral-8b†	–	mistralai/Ministral-3-8B-Base-2512	b
weihongliang/jevtrainer-qwen35-0.8b-2026-10-02†	–	Qwen/Qwen3.5-0.8B	c
knowlet/Gemma-4-12B-Unified-System-One†	–	google/gemma-4-12B-it	c
decision-tune/decisiontune-1.0†	–	answerdotai/ModernBERT-large	b
ParallaxOpen/Vela-Decision-170M†	–	–	b
yuanphd/jevmark†	–	Qwen/Qwen3-0.6B-Base	c

Repository	Parameters	Base model	Head
theshekslaw/clinical-jev-qwen3.5-2b-mtsamples†	–	Qwen/Qwen3.5-2B-Base	c
chendren/cx-decider-applied-v1†	–	Qwen/Qwen3.5-2B-Base	b
weihongliang/jevtrainer-qwen35-2b-2026-10-05†	–	Qwen/Qwen3.5-2B	c
perplexity-ai/pplx-decider-v1.1-27b†	–	Qwen/Qwen3.8-27B	a
gabrielfior/decider-q2b-1.1b†	–	Qwen/Qwen3.5-2B-Base	b
ZhangPY/kev-snake-vision-grpo†	–	Qwen/Qwen3.5-4B-Base	b
AlephFunk/jev-qwen†	–	Qwen/Qwen3.5-4B-Base	c
StrandsAgents/strands-decider-2B-hobson-v21†	–	Qwen/Qwen3.5-2B-Base	b
cribl-ai/cribl-decision-1.0†	–	Qwen/Qwen3.5-4B-Base	c
sgaseretto/diffcider-typed-decisions†	–	dllm-hub/Qwen3-0.6B-diffusion-mdlm-v0.1	c
shunmoridev/llm-jp-4.1-33b-systemone-experimental†	–	llm-jp/llm-jp-4.1-33b-thinking	c
shgao/rsi-jev-v6.0-vl-4b†	–	Qwen/Qwen3.5-4B-Base	b
systemonedev/kenning-large-v0.5†	–	MoritzLaurer/deberta-v3-large-zeroshot-v2.0-c	d
thinkingdbx/DECISIONx-1.5B†	–	Qwen/Qwen2.5-1.5B-Instruct	b
odunola/lfm2.5_jev_style†	–	LiquidAI/LFM2.5-350M	b
systemonedev/kenning-xl-v0.6†	–	Qwen/Qwen3-1.7B	c
shgao/rsi-jev-v6.1-vl-4b†	–	Qwen/Qwen3.5-4B-Base	b
LMDHQ-0420/vision-jev†	–	Qwen/Qwen3.5-0.8B	b
UlkuTuncerKucuktas/vistalab-system-one-12b†	–	–	c
UlkuTuncerKucuktas/vistalab-system-one-e4b†	–	google/gemma-4-E4B-it	c
Mapika/decider-31b†	–	google/gemma-4-31B-it	c
j-raghavan/s1-gemma4-26b-decision†	–	google/gemma-4-26B-A4B-it	c
chenz53/Jev-alpha-26B-A4B†	–	google/gemma-4-26B-A4B-it	c
ksonny4/modernbert-jev-v2†	–	–	d
shgao/rsi-jev-v6.1-vl-27b†	–	Qwen/Qwen3.8-27B	b
systemonedev/kenning-large-v0.6†	–	tasksource/ModernBERT-large-nli	d
weihongliang/jevtrainer-qwen35-2b-agent-2026-10-08†	–	Qwen/Qwen3.5-2B	c
HanWangTJU/haidass1.5-143m-jev†	–	DALabCommunity/Haidass1.5-143M	b
Maincode/Matilda-Jev-v1.5†	–	Qwen/Qwen3.8-27B	a
Cloudflare/clef-omni†	–	Qwen/Qwen3-Omni-30B-A3B-Instruct	b
StrandsAgents/strands-decider-E2B-gemma4-v1-2610†	–	google/gemma-4-E2B-it	b
StrandsAgents/strands-decider-12B-gemma4-v1-2610†	–	google/gemma-4-12B-it	b
StrandsAgents/strands-decider-E4B-gemma4-v1-2610†	–	google/gemma-4-E4B-it	b
StrandsAgents/strands-decider-26B-A4B-gemma4-v1-2610†	–	google/gemma-4-26B-A4B-it	b
cribl-ai/cribl-decision-1.0-0.8B†	–	Qwen/Qwen3.5-0.8B	c
StrandsAgents/strands-decider-2B-qwen3.5-v1-2610†	–	Qwen/Qwen3.5-2B-Base	b
lxls/typed-decider-150m-v1†	–	jhu-clsp/ettin-encoder-150m	b
bespokelabs/Bespoke-Nimble-27B-v3†	–	–	u
SargeDev/Solar-JEV-Gemma-embed-2†	–	google/embeddinggemma-2	a
leinenterprises/jev-vision-27b†	–	Qwen/Qwen3.8-27B	c
AstitwaAgarwal/multimodal-jev†	–	Qwen/Qwen3-VL-2B-Instruct	b

Repository	Parameters	Base model	Head
startlux-models/StartLux-Decision-v1.1-0.8B-Instant†	–	–	c
startlux-models/StartLux-Decision-v1.1-2B-Instant†	–	–	c
startlux-models/StartLux-Decision-v1.1-4B-Instant†	–	–	c
startlux-models/StartLux-Decision-v1.1-4B-AutoThink†	–	–	c
startlux-models/StartLux-Decision-v1.1-9B-Instant†	–	–	c
startlux-models/StartLux-Decision-v1.1-9B-AutoThink†	–	–	c
startlux-models/StartLux-Decision-v1.1-27B-AutoThink†	–	–	c
wozniakos10/Qwen3.5-4B-typed-decisions†	–	unsloth/Qwen3.5-4B	b
autotrust/JEV-35B†	–	Qwen/Qwen3.5-35B-A3B	a
startlux-models/StartLux-Decision-v1.1-27B-Instant†	–	–	c
Jahirrrr/Klyra-Decision-16M†	–	–	d
flymy-ai/decision-4b-v1.6†	–	Qwen/Qwen3.5-4B	c
J33-AI/j33-jev†	–	google/embeddinggemma-2	b
Nikunjgoyal07/jev-qwen-decider†	–	Qwen/Qwen3-0.6B	b
Nikunjgoyal07/jev-deberta-decider†	–	microsoft/deberta-v3-base	b
Ruivalim/exu-qwen3.5-2b-typed-decisions†	–	Qwen/Qwen3.5-2B	d
workmagicio/workmagic-jev-decision†	–	Qwen/Qwen3-4B	c
Louiskq/jev-class-9b-sft†	–	Qwen/Qwen3.5-9B-Base	c
thealper2/qwen3.5-0.8b-open-jev-decision†	–	unsloth/Qwen3.5-0.8B	b
TurkishCodeMan/KUTUP-1-Decision-2B†	–	Qwen/Qwen3.5-2B-Base	b
joebumodels/JevFinetune†	–	Qwen/Qwen3.5-4B	b
faizaltkl/qdec-lay-decision-model†	–	unsloth/Qwen3.5-4B	b
Skyl4r-Ai/Skylar-980M-Cobol-Decider†	–	Skyl4r-Ai/Skylar-980M-Cobol-Base	b
mrm8488/embeddinggemma-2-ft-typed-decision†	–	google/embeddinggemma-2	d
Steerix-org/robo-jev-2b-lora†	–	Qwen/Qwen3.5-2B	c
haohan84/v-jev†	–	Qwen/Qwen3.5-4B	c
ejbejaranos/JEV-HRI-Qwen3-VL-2B†	–	Qwen/Qwen3-VL-2B-Instruct	c
ejbejaranos/JEV-HRI-Qwen3-VL-4B†	–	Qwen/Qwen3-VL-4B-Instruct	c
ejbejaranos/JEV-HRI-Qwen3-VL-8B†	–	Qwen/Qwen3-VL-8B-Instruct	c
rony/gemma3-270m-typed-decisions†	–	google/gemma-3-270m-it	c
myeongcheol/decision-lfm2.5-1.2b†	–	LiquidAI/LFM2.5-1.2B-Instruct	c
siriusyiyi/my_jev-hr9†	–	Qwen/Qwen3-0.6B-Base	c
ejbejaranos/JEV-HRI-SmolVLM2-256M†	–	HuggingFaceTB/SmolVLM2-256M-Video-Instruct	c
ejbejaranos/JEV-HRI-LFM2-VL-450M†	–	LiquidAI/LFM2-VL-450M	c
workmagicio/workmagic-jev-churn-distill-qwen3-8b-v1†	–	Qwen/Qwen3-8B	c
Singularity-Compute/super-decider†	–	google/diffusiongemma-26B-A4B-it	c
johannhartmann/gemma-4-E4B-it-unsloth-bnb-4bit_qlora_typed-decisions-de†	–	google/gemma-4-E4B-it	b
itsdevice/xlm-roberta-systemone†	–	FacebookAI/xlm-roberta-base	b

Repository	Parameters	Base model	Head
rony/ministral3-3b-reasoning-typed-decisions†	–	mistralai/Ministral-3-3B-Reasoning-2512	c
berkay123001/turnike-jev-4b†	–	Qwen/Qwen3.5-4B	c
berkay123001/turnike-jev-0.8b†	–	Qwen/Qwen3.5-0.8B	c
berkay123001/turnike-jev-9b†	–	Qwen/Qwen3.5-9B	c
zhangjiaxin2012/SRF-Jev-Qwen3.5-4B-Historical-8876†	–	Qwen/Qwen3.5-4B	c
xmutly/OmniMed-Jev-4B†	–	google/medgemma-1.5-4b-it	b
Danny-Dasilva/embeddinggemma-2-multimodal-typed-decision†	–	google/embeddinggemma-2	b
assix-research/jev-system1-gemma4-26b-decide†	–	google/gemma-4-26B-A4B-it	b
Maincode/Matilda-Decision-v1†	–	Qwen/Qwen3.8-27B	a
perplexity-ai/pplx-decider-v1.2†	–	Qwen/Qwen3.8-27B	a
j-llm/jevlite†	–	answerdotai/ModernBERT-base	b
decisionmodels/s1-llm-auto-router†	–	jhu-clsp/mmBERT-small	a
Nefelim584/jev-lite-stage-a†	–	answerdotai/ModernBERT-large	b
SlayerLab/Slayer-149M-Decision-v1†	–	–	c
perplexity-ai/pplx-decider-flash-v1.2†	–	Qwen/Qwen3.5-4B	a
JevAgentHarness/AgentDecision-4B†	–	Qwen/Qwen3-4B	c
JevAgentHarness/AgentDecision-4B-v1†	–	Qwen/Qwen3-4B	c
Rology/rad-decider†	–	Qwen/Qwen3.5-0.8B	c
MachineFi/Trio-Spark†	–	–	u
gyanbhambhani/oss-jev-qwen05-decision-10k†	–	Qwen/Qwen2.5-0.5B-Instruct	c
gyanbhambhani/oss-jev-qwen05-decision-noul-bin†	–	Qwen/Qwen2.5-0.5B-Instruct	c
shibing624/train-jev-4b-lora†	–	Qwen/Qwen3.5-4B	a
shibing624/train-jev-27b-lora†	–	Qwen/Qwen3.8-27B	a
wiiiktor/bielik-minitron-jev-pl†	–	speakleash/Bielik-Minitron-7B-v3.0-Instruct	b
modelfaces/granite-4.1-30b-decision†	–	ibm-granite/granite-4.1-30b-base	b
Elfsong/jev-flan-t5-large-20261017-public-v2†	–	google/flan-t5-large	b
SnowGemma/Qwen3.5-0.8B-Typed-Decisions†	–	Qwen/Qwen3.5-0.8B	b
sp1d5r/video-jev†	–	Qwen/Qwen3-Omni-30B-A3B-Instruct	b
Elythros/elythros-decider-0.6b†	–	Qwen/Qwen3-0.6B	b
bespokelabs/Nimble-v4-max†	–	Qwen/Qwen3.8-27B	c
cribl-ai/cribl-decision-1.0-9B†	–	Qwen/Qwen3.5-9B	c
ronith128/969M-Decisions-CPU†	–	LiquidAI/LFM2.5-1.2B-Instruct	c
cribl-ai/cribl-decision-1.0-27B†	–	Qwen/Qwen3.5-27B	c
scairesearch/kieve-jev†	–	–	d
bizhmans/krasivo-jev-no-bench-bf16†	–	Qwen/Qwen3.5-2B-Base	b
zhangjiaxin2012/SRF-Jev-Gemma4-12B-LoRA†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-26B-A4B-LoRA†	–	google/gemma-4-26B-A4B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-31B-LoRA†	–	google/gemma-4-31B-it	c
bizhmans/krasivo-jev-bf16†	–	Qwen/Qwen3.5-2B-Base	b
scairesearch/kieve-jev-100m†	–	–	d
NeelContractor/jev-decider-0.4B†	–	Qwen/Qwen3-0.6B	b
Mapika/decider-12b-vision†	–	google/gemma-4-12B-it	c
MaziyarPanahi/Qwen3.5-4B-Jev-Judge†	–	Qwen/Qwen3.5-4B	b
HallD/Yev-Decision-1†	–	Qwen/Qwen3.5-4B-Base	b

Repository	Parameters	Base model	Head
Futurix-AI/FxAI-decision-model†	–	–	a
SargeDev/Solar-JEV-27B†	–	huihui-ai/Huihui-Qwen3.8-27B-abliterated	a
Elythros/elythros-decider-1.7b†	–	Qwen/Qwen3-1.7B	b
scairesearch/kieve-jev-large†	–	–	d
scairesearch/kieve-jev-v0.7†	–	–	d
tensorninja2/decision-1†	–	Qwen/Qwen3.8-27B	b
turion-ai/turion-jev-31b†	–	google/gemma-4-31B-it	c
IntelligenceLab/Jev-Vision-0.8B†	–	Qwen/Qwen3.5-0.8B	c
IntelligenceLab/Jev-Vision-2B†	–	Qwen/Qwen3.5-2B	c
forward-metric/sysone-typed-decisions-0.4B†	–	jhu-clsp/ettin-encoder-400m	b
jmrenouard/microllm-rlcd2-decision-head†	–	–	d
0xBrandon/Decision-WAF-V1-0.8B†	–	Qwen/Qwen3.5-0.8B-Base	c
Jahirrrr/Klyra-Decision-2B-V1†	–	openbmb/MiniCPM5-2B-Midtrain	d
Elfsong/jev-ettin-150m-v16†	–	cross-encoder/ettin-reranker-150m-v1	b
zhangjiaxin2012/SRF-Jev-Gemma4-12B-Stage1-SFT1024-Think4K-v2†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-31B-Stage2-Causal-LoRA1024-Think4K-v2†	–	google/gemma-4-31B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-31B-Stage1-SFT1024-Think4K-v2†	–	google/gemma-4-31B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-12B-Stage2-Causal-LoRA4096-Think4K-v2†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-12B-Stage3-Legacy-Reasoning-SFT128-Think4K-v2†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-12B-Stage3-Legacy-LoRA-RL16-Train4K-Think4K-v2†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-31B-Stage3-Reasoning-SFT128-Think4K-v2†	–	google/gemma-4-31B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-31B-Stage3-Reasoning-SFT1024-Think4K-v2†	–	google/gemma-4-31B-it	c
sparrow-decider/sparrow-nano-general†	–	minishlab/potion-base-8M	d
sparrow-decider/sparrow-nano-robot†	–	minishlab/potion-base-8M	d
sparrow-decider/sparrow-tiny-general†	–	jhu-clsp/ettin-encoder-32m	b
sparrow-decider/sparrow-small-general†	–	jhu-clsp/ettin-encoder-68m	b
zhangjiaxin2012/SRF-Jev-Gemma4-26B-A4B-Stage2-Causal-LoRA1024-Think4K-v2†	–	google/gemma-4-26B-A4B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-12B-Stage3-Full-RLCD2-16-Train4K-Think4K-v2†	–	google/gemma-4-12B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-26B-A4B-Stage3-Reasoning-SFT128-Think4K-v2†	–	google/gemma-4-26B-A4B-it	c
zhangjiaxin2012/SRF-Jev-Gemma4-26B-A4B-Stage3-Reasoning-SFT1024-Think4K-v2†	–	google/gemma-4-26B-A4B-it	c
0xBrandon/Decision-WAF-V1.1-0.8B†	–	Qwen/Qwen3.5-0.8B-Base	c
0xBrandon/Decision-WAF-V1-0.1B†	–	AxiomicLabs/GPT-X3-150M	c
scairesearch/kieve-jev-moe-300m†	–	–	d
scairesearch/kieve-jev-fractal-300m†	–	–	d
scairesearch/kieve-jev-curriculum-300m†	–	–	d
q-project/N-Q-26M-JEV†	–	q-project/N-Q-26M-Chat	b
Maincode/Matilda-Decision-v1-Edge†	–	Qwen/Qwen3.5-0.8B	a
chaoliangUNSW/Jev-Style-2B-Decision-v4†	–	Qwen/Qwen3.5-2B	c

Table S8. T2 derivatives by parent T1 or T2 checkpoint (Section 6), for every parent with at least five recorded derivatives. The last row counts the other parents and their derivatives; the full list is in the companion repository.

Parent	Derivatives	Parent	Derivatives
convaiinnovations/laya	386	perplexity-ai/pplx-decider-flash-v1.2	10
convaiinnovations/laya-multilingual	113	StrandsAgents/strands-decider-2B-hobson-v21	8
Cloudflare/clef-flash	78	autotrust/JEV-27B	8
Cloudflare/clef	62	chaoliangUNSW/Jev-Style-0.8B-Decision-v3	7
jaredpalmer/kev-4b	48	jaredpalmer/kev-0.6b	7
jaredpalmer/kev-0.8b	35	akhilaaa3/Jev-Omni	6
convaiinnovations/laya-typed-decisions	31	autotrust/JEV-35B	6
Mapika/decider-2b	28	jaredpalmer/kev-27b	6
Mapika/decider-4b	19	pashwinkrish22/laya-email-resume-classifier-v3	6
Terom/laya	17	perplexity-ai/pplx-decider-v1.2	6
jaredpalmer/kev-9b	17	suryatmodulus/clef	6
Cloudflare/clef-omni	16	Mapika/decider-2b-vision	5
Mapika/decider-0.8b	16	StrandsAgents/strands-decider-2B-hobson-v19	5
iapp/OpenThai-SystemOne	16	StrandsAgents/strands-decider-E4B-gemma4-v1-2610	5
autotrust/JEV-27B-VL	15	TokenRhythm/NeoHorse-Jev-4B	5
perplexity-ai/pplx-decider-v1.1-27b	12	jaredpalmer/kev-0.5b	5
alibiserikbay/JevK5	10	pashwinkrish22/laya-email-resume-classifier	5
autotrust/JEV-9B	10	startlux-models/StartLux-Decision-v1-0.8B	5
Other parents (219), each with fewer derivatives			330

Repository	Parameters	Base model	Head
Sagicc/nanoGentzen-Jev-v3†	–	–	d

S6 Pattern catalog entries

Each entry expands a row of Table 5 in Section 9. That table already states the quantity each pattern consumes, the status of its guarantee and the failure modes of Table 6 it mitigates, so the entries give only structure and evidence.

The gate makes one call and acts automatically only when the returned probability clears a threshold declared in code. Rafe and Das [179] names conformal risk control, which gives a finite-sample guarantee on exchangeable data [3, 10], as the stronger alternative to its own held-out estimate. In Ibrahim and Zaki [80] the median task at a fixed confidence threshold covered 0.376 of items at an accuracy of 0.815, and Li et al. [116] accept judgments only above a confidence threshold.

Screen-then-read runs an inexpensive schema over every record and the full schema only over forwarded records. By the recall identity of Proposition 5.4, final recall is the product of the measured rates at which the screen forwards positives and the reader detects them, so an evaluation on

forwarded records alone reports only the second. Rafe and Das [179] ran a lean presence schema over a random frame of narratives and the full schema over the forwarded records.

Candidate-then-select has a retriever, rule or other model propose a short list for the decision model to choose from, which keeps the answer space small because Proposition 5.3 shows that one threshold admits more risk as that space grows. The list ideally includes a none-of-the-above option, which the vendor recommends when a list may not cover every input [225], and final accuracy is bounded by the recall of the list. The decider-2b card reports lower accuracy with full than with sampled label sets [147], and REFLEX lost −6.7 accuracy points as options grew [254].

Speculative fan-out asks many questions about one state in one call, some possibly unneeded, and lets code pick the answers, relying on the shared state encoding of Section 4. The vendor claims cost and speed benefits [228], but Proposition 5.1 identifies a conjunction only up to an interval, so multiplying the answers asserts an independence that the contract does not supply. An independent comparison found hosted latency flat in question count [9], and a visual decision model ran faster by a factor of 8.9 with shared-prefix batching [272].

The confidence-gated cascade answers with the decision model when its top probability clears a threshold and escalates the rest to a frontier model, to retain most frontier quality at a fraction of its fee. In 23 studies escalation kept 0.67 to 1.2 of frontier quality at 0.18 to 0.92 of its fee under differing cost definitions [80, 116, 254]. The saving shrank as the price gap narrowed, and with already accurate routing a cheap generative cascade matched the decision-model cascade at 0.72 of its cost [254].

Judge-and-escalate applies the same logic to judging another system's output, such as a pairwise preference or the factual support of a statement [284], and escalates unconfident judgments to a stronger judge or a person. Li et al. [116] found 47.7 percent accuracy at low confidence against 95.8 percent at high confidence, so confidence ordered the errors, and they recommend escalation on difficult correctness and on style-adversarial pairs. Huang et al. [76] judged radiology-report factuality with one support question per statement.

Composite scoring asks each dimension of an assessment as its own question and combines the answers by a fixed rule in code. The vendor credits it with cost, reliability and speed benefits [228] and advises against bundling judgments into one question [226]. The combined score is calibrated only if the rule is fitted, Proposition 5.1 limits reading the rule as a joint probability, and Proposition 5.6 warns against feeding it expected levels. Robitza [188] compared a fitted linear rule over decomposed outputs with trained baselines on the same features.

Hierarchical fallback replaces one Choice over many answers with a chain asking a coarse category and then an item, falling back to the category when the fine stage is unsure. By the stage-product accounting of Proposition 5.4, a chain can be no more accurate than its coarse stage when a correct item implies a correct category. REFLEX measured accuracy loss with a growing option count [254], and the companion benchmark compares one Choice over all CLINC-150 intents with a domain-then-intent chain [178].

Paraphrase self-consistency asks one decision in several equivalent forms that vary wording, option order or option names, and flags for review any disagreement, read from the spread of the distributions alongside their average. Exchangeable random names implement the option-name neutralization of Proposition 5.5, giving expected log-odds that depend on the rubrics alone, and swaps with fixed rubrics identify the name term. Opaque random names flipped 2.02 answers per hundred on the hosted model, in the neutral regime [212]. A held-out paraphrase flipped 5.2 percent of JevLite decisions and an option-order swap 2.6 percent [186], and Zhang et al. [276] compared output orders across models.

Deterministic pre-parsing computes in code whatever code computes exactly, such as counts, date differences and unit conversions, and passes the result to the model inside the state. This follows

the vendor's remedies of keeping arithmetic in code, extracting and comparing date components in code, and reducing hops of indirection [226]. The Kev-9B card reports deadline accuracy of 0.80 (raw) / 0.90 (with date_facts preprocessor) against 0.93 for the hosted model [163], and the preprocessor that supplies date facts to the state accounts for the difference.

Marginal-to-joint reconciliation serves programs that need a statement about several answers together, which question isolation does not supply. It asks one Choice over the joint outcomes, reports an interval for the conjunction, or enforces logical identities across related answers in code, in line with the vendor's advice to put each decision one way [226]. The bounds of Proposition 5.1 are attained, so the interval is the most that the marginals support. In the vendor's worked example a Noul and a Choice about one fact disagree [226], and Cheng et al. [32] report parity judgments consistent with related band judgments in 0.513 of cases.

The review budget sends to human review the records whose calibrated positive probability falls in the band where review costs less than the expected loss of automation. The budget is the probability mass inside the band of Equation (7). Rafe and Das [179] use a precision-targeted deferral variant, not a cascade result, in which a target precision among flagged records sets the upper edge, with a pooled held-out budget of 0.8 percent of flagged records. Their in-domain recalibration moved pooled calibration error from 0.0231 to 0.0069.

The versioned contract pins the model version, schema, option names and rubrics, and serving temperature, which jointly define the predictor (Section 4), and repeats the calibration audit when any of them changes. The hosted service separates pinned versions from moving aliases [225], and pinning makes a change detectable but not preventable [179]. In the motivating census case, Bespoke-Nimble-9B moved from a checkpoint at a fitted temperature with fewer choices per field to a revision at its default temperature with more choices and a warning against reusing the earlier fit [13, 14]. The repository's endpoints disagree on the earlier tag, and the companion benchmark loads the current revision while the ledger rows on Nimble concern the earlier one, so the two describe different predictors. Kev-0.8B also changed serving temperature between checkpoints under an unchanged name [162].

S7 The evidence ledger by study

Table S9 lists the evidence ledger of Section 7 one study at a time, including the grey-literature sources under their bibliography keys, so that a reader can check a study before reusing its numbers. Its Models column counts the distinct values of the ledger's model field and can undercount comparators recorded in a separate field. Its n is the most frequent numeric sample size across a study's rows, or else the largest leading number among its descriptive entries, with two values corrected by hand against each study's stated primary sample. Table S10 regroups a subset of the studies by the seven application areas of Section 10. Most studies fall in one of these areas, and the rest are one classification-robustness study, one reverse-engineering probe of the hosted service, one community-built benchmark and four open decision models each compared against Jev, several of them cited in Section 6. Of the 248 studies, 3 received an overall grade of low risk of bias. Reference labels often mix human adjudication, an automated judge and a computed task outcome rather than one uniform gold standard.

Table S9. Every study in the evidence ledger, by task family, its overall risk-of-bias grade, its reference-label type and the number of ledger rows it contributes, with Models and n counted as Section S7 describes.

Study	Task family	Models	n	Reference labels	Rows	RoB
Ahmed [1]	classification	7	240	administrative field; none	12	high
Amit [2]	other	7	5,000	not reported	17	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Antunes et al. [4]	classification	1	9,092	human crowd; human adjudicated; synthetic; llm teacher; none	13	some concerns
Asseri and Asseri [5]	other	1	36,000	none; human crowd	13	high
Azizi et al. [6]	classification	11	833	human adjudicated; synthetic	16	some concerns
Azizi et al. [7]	safety	7	1,034	human single; synthetic	19	high
Bailo [8]	agent control	6	600	synthetic	14	high
Bellver-Soler et al. [11]	classification	1	3,113	human crowd; none	15	high
Besanson [12]	extraction	1	700	synthetic	23	high
Brom [16]	classification	3	1,000	human crowd; none; task outcome	17	high
Bru et al. [17]	other	4	9,621	human crowd; none	9	some concerns
Buecheler et al. [18]	agent control	1	144	none	11	some concerns
Bukhari et al. [19]	classification	4	231	administrative field; synthetic; human single; none; human crowd; task outcome; llm teacher	15	some concerns
Cao et al. [21]	retrieval	5	100	administrative field; task outcome	18	high
Chen [22]	safety	8	607	human adjudicated; none; administrative field	12	some concerns
Chen [24]	retrieval	7	276	human single	17	some concerns
Chen and Li [23]	other	2	1,000	synthetic	17	high
Chen et al. [27]	QA; classification	1	not reported	not reported; human crowd	14	high
Chen [25]	agent control	5	268	administrative field; human single; none	21	some concerns
Chen et al. [26]	classification	8	8,000	human single	12	high
Chen et al. [30]	safety	6	1,000	administrative field; synthetic; none; human adjudicated	21	some concerns
Chen and Le [29]	memory	3	500	model consensus; human single	15	some concerns
Chen et al. [28]	other	7	78,362	mixed; none; task outcome	14	some concerns
Cheng et al. [31]	other	1	72	not reported	9	high
Cheng et al. [32]	judging	1	68	task outcome	12	high
Chouhan and Perry [33]	classification; QA	5	4,689	llm teacher; human single; human crowd; none	17	high
Chung et al. [35]	classification	5	11,117	human adjudicated; synthetic	16	high
Cinà [36]	other	2	978	task outcome; none	10	high
Dai et al. [40]	classification	4	353	human crowd; administrative field; synthetic; human single	10	some concerns
Dai et al. [41]	other	3	36,000	synthetic	6	high
Deka [42]	classification	16	2,000	human adjudicated; none	20	some concerns

Study	Task family	Models	*n*	Reference labels	Rows	RoB
Demir [43]	QA	6	5,314	other: court-stated holding from the citing opinion (CaseHOLD gold); abstain condition built programmatically	14	some concerns
Deng et al. [44]	scientific workflow	7	100	human adjudicated	14	some concerns
Denney and DiGiuseppe [45]	social-science annotation; classification	1	37,709	human adjudicated; human single; human crowd; none	23	high
Deußer et al. [46]	QA; classification; extraction; safety; judging; other; multi-dataset zero-shot evaluation	1	755	human single; human crowd; administrative field; none	26	some concerns
Ding et al. [47]	classification; judging; other	2	350	not reported; human crowd; human single; llm teacher; synthetic; none	17	some concerns
Ding [48]	QA	7	14,042	human adjudicated	11	high
Dinzinger et al. [49]	retrieval	12	not reported	human single	17	high
Dong et al. [50]	safety	1	12	synthetic	9	high
Rosario and Atef [189]	judging	5	495	human adjudicated	18	high
dos Santos Barbosa [51]	agent control	2	13	task outcome; none; vendor label	14	high
Du et al. [52]	classification	2	8,742	human adjudicated	17	high
Duan et al. [53]	other	11	100	administrative field; synthetic; task outcome; llm teacher; human crowd	20	high
Fadli et al. [56]	classification	6	4,800	human single	21	high
Feng et al. [57]	other	3	9,845	human single; task outcome	13	high
Fernandes and Pereira [58]	judging; retrieval	1	8,205	llm teacher; human adjudicated; human single	14	some concerns
Gan et al. [60]	classification	9	10,027	human crowd; synthetic; model consensus	17	high
Gao et al. [61]	social-science annotation; classification	5	180,000	human crowd	17	high
Gershon et al. [62]	judging	1	118	other: physician panel (median of five physicians); other: physician labels (2 to 5 of 186 physicians per row); other: clinician ratings (56 responses) and LLM jury scores; llm teacher; other: mixed (physician panel, clinicians, LLM graders); other: physician labels (UC2) and GPT-5.4 (UC4)	16	some concerns

Study	Task family	Models	*n*	Reference labels	Rows	RoB
Geuter et al. [63]	classification	4	2,352	model consensus; administrative field; human adjudicated; none	16	high
Hume, Archer [77]	QA; other; classification	1	1,200	human adjudicated	8	some concerns
Bespoke Labs [14]	classification	3	324	synthetic; human adjudicated; human crowd	10	some concerns
Palmer, Jared [163]	classification; QA; other	4	764	human adjudicated; administrative field; none	12	high
Convai Innovations [39]	classification	6	2,000	llm teacher; human adjudicated	10	some concerns
LocalLLaMA [132]	classification	6	2,000	llm teacher	10	high
Balian, Rob [9]	classification	6	975	human adjudicated; task outcome	8	some concerns
TypeSafe AI [230]	agent control	5	not reported	model consensus	9	high
Rauch [185]	safety	1	not reported	none	1	high
Guo et al. [66]	judging	1	1,228	human adjudicated; synthetic	10	high
Guo et al. [65]	safety	1	7,193	llm teacher; human adjudicated; none	20	some concerns
Ha et al. [67]	classification	4	893	human single	18	high
Han et al. [69]	agent control	3	100	task outcome	6	some concerns
Han et al. [68]	agent control	6	3,875	administrative field	13	high
He et al. [70]	classification	1	322	other: source-derived reference labels (TrialGPT annotations; MIMIC questions model-generated and human-reviewed); none	13	some concerns
Hu et al. [72]	safety	7	5,586	model consensus; task outcome	19	high
Hu et al. [73]	safety; agent control	2	7,033	administrative field; none; task outcome; synthetic	13	some concerns
Huang et al. [76]	judging	2	100	human adjudicated; task outcome; none	13	some concerns
Huang et al. [75]	classification; agent control	11	316	human adjudicated; none; task outcome	14	high
Huang et al. [74]	safety	3	1,947	administrative field; none	13	some concerns
Huo et al. [78]	other	3	43,496	synthetic; not reported	13	high
Ibrahim and Zaki [80]	social-science annotation	27	7,977	human adjudicated	55	high
Ichikawa [81]	other: multi-task decision benchmark	5	600	other: benchmark labels aggregated in S1MB (mixed sources); none	10	some concerns
Imperatrice et al. [82]	other	1	25,000	task outcome; none	17	low
Iqbal et al. [83]	classification	7	970	human crowd	20	some concerns

Study	Task family	Models	*n*	Reference labels	Rows	RoB
Jiang et al. [84]	memory	2	not reported	llm teacher	10	high
Joy [85]	classification	2	1,000	human crowd; human adjudicated	21	high
Jung [86]	classification	1	128	task outcome	16	some concerns
Khatri et al. [87]	classification	5	406	other: expert annotator labels (adjudication not stated)	12	high
Khezresmaeilzadeh et al. [88]	agent control	3	134	task outcome	15	some concerns
Konak [89]	classification	2	890	administrative field	19	some concerns
Kourtis and Xilouris [91]	agent control	5	3,645	synthetic; none	13	high
Kwartler et al. [92]	classification	1	109	human single; none	10	some concerns
Lai et al. [93]	other	2	384	synthetic; human single	15	some concerns
Lai et al. [94]	other: executable arithmetic rule following	1	320	synthetic	8	high
Lan et al. [96]	safety	4	1,332	human adjudicated	10	some concerns
Lan et al. [95]	safety	1	47,611	synthetic; none	15	high
Lee et al. [97]	QA	1	1,273	human adjudicated; task outcome; none	16	high
Lee and Park [98]	other	5	14,000	task outcome	14	some concerns
Li [111]	classification	2	22,162	human adjudicated	11	high
Li et al. [106]	other	3	3,360	none; human crowd	15	some concerns
Li et al. [114]	classification	1	2,584	human single	11	high
Li and Xie [108]	retrieval	13	443	other: benchmark relevance judgments (incomplete; 607 of 1,384 BRIGHT candidate sets hold a judged positive)	16	some concerns
Li et al. [109]	classification	5	1,120	human single	9	some concerns
Li et al. [101]	agent control	2	63	task outcome; administrative field; human adjudicated	12	some concerns
Li [105]	agent control	4	2,100	human adjudicated; task outcome; none; llm teacher	19	high
Li and Wang [107]	classification	2	3,080	llm teacher; human adjudicated; none	13	high
Li et al. [116]	judging	3	400	vendor label; none; human adjudicated	27	some concerns
Li and Chen [110]	classification	10	500	none; human crowd; human single	15	high
Li [104]	social-science annotation; classification; other	8	9,070	human crowd	10	high
Li and Wagle [117]	classification; QA	6	231	not reported; human adjudicated	11	some concerns
Li et al. [115]	agent control	7	65,228	task outcome; synthetic	21	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Li et al. [102]	agent control	3	216	task outcome; llm teacher; human adjudicated	12	some concerns
Li et al. [118]	social-science annotation	9	240,458	human single; none	10	some concerns
Li et al. [113]	safety	1	200	human adjudicated; none	14	high
Li [103]	agent control; QA; retrieval	12	6,396	human single; not reported	17	high
Li et al. [100]	agent control	1	120	synthetic	20	high
Li et al. [112]	QA	1	1,172	human single	12	high
Li et al. [99]	other	1	936	llm teacher; human single	8	high
Lin [119]	classification	4	324	synthetic	15	high
Ling et al. [120]	ecosystem analysis of public projects	1	not reported	none	1	not graded
Ling et al. [121]	agent control	1	348	task outcome	13	some concerns
Liu [127]	judging	6	200	synthetic	13	high
Liu et al. [124]	other: sequential recommendation	2	1,280	task outcome	18	high
Liu et al. [128]	safety	5	9,852	other: mixed benchmark annotations (human span labels, GPT-4o-injected errors validated by two annotators, constraint checks); none	16	high
Liu [129]	safety	7	236	administrative field	22	high
Liu et al. [130]	retrieval	3	100	administrative field; none	8	high
Liu et al. [126]	social-science annotation	5	14,412	other: mixed source-dataset annotations (gold and silver, human); none; other: public dataset labels (construction not described)	12	high
Liu et al. [122]	classification	2	4,625	model consensus; synthetic; none	10	high
Liu et al. [125]	agent control	1	5	task outcome	8	high
Liu et al. [123]	classification	9	13,000	administrative field	10	high
Loi [133]	judging	1	64	none	8	high
Lu and Wu [134]	classification	4	166	administrative field	15	high
Lucca [135]	classification	6	722	human single	16	some concerns
Lukassen et al. [136]	judging	3	439	administrative field; synthetic; none	15	high
Lukauskas and Šarkauskaitė [137]	social-science annotation	9	100	human crowd	20	high
Luo et al. [138]	safety	1	2,273	none; llm teacher	7	some concerns
Lyu and Xia [139]	retrieval; other	2	1,000	task outcome	14	high
Ma et al. [144]	agent control	1	50	task outcome	8	some concerns
Ma et al. [140]	other	10	1,504	human single; synthetic	11	some concerns

Study	Task family	Models	n	Reference labels	Rows	RoB
Ma et al. [142]	agent control	5	105	task outcome	10	some concerns
Ma et al. [141]	judging	2	2,334	task outcome; synthetic	21	high
Ma et al. [143]	other: constrained spatial planning for 3D world generation	1	100	other: programmatic hard-constraint check (terrain, water, biome, Macro role, adjacency)	2	some concerns
Madrid-García and Merino-Barbancho [146]	QA	1	1,373	human adjudicated	20	some concerns
Mazzoni and Failla [148]	safety	2	3,689	administrative field; human crowd	17	high
Mei et al. [149]	QA	8	231	administrative field; none	9	some concerns
Messoudi and Mansar [150]	moderation; safety; retrieval	3	1,050	administrative field; none	19	high
Min and Chen [153]	judging; agent control	2	300	llm teacher; task outcome; none	16	high
Mohebifar and Bit [155]	other	8	1,932	synthetic; human adjudicated; task outcome; human single	17	high
Mu et al. [156]	classification	4	1,279	synthetic; task outcome	14	high
Munot [158]	other: multi-correct-choice probability allocation	3	7,405	other: dataset gold labels as released	16	high
Nandakishore [159]	judging	2	not reported	human crowd; task outcome; none	17	high
Nandakishore [160]	agent control	3	2,000	llm teacher; none; human single	15	high
Pan et al. [164]	agent control	11	60	task outcome	14	some concerns
Panchal [165]	QA; agent control	5	12,032	human adjudicated; none; task outcome	13	some concerns
Paramasivam and Rajkumar [166]	judging	2	235	llm teacher; human crowd; none	11	some concerns
Paschalides et al. [167]	safety	7	4,120	human crowd; synthetic	17	high
Pei et al. [168]	agent control	6	1,011	task outcome	10	some concerns
Pelaseyed and Zabihi [169]	other	4	96	none	11	some concerns
Perifanis et al. [170]	other	5	11,481	task outcome	14	some concerns
Polignano et al. [173]	other	13	2,000	task outcome	18	some concerns
Porcedda [174]	other	1	365	synthetic	13	high
Qian et al. [175]	agent control	1	1,168	administrative field	8	high
Qin et al. [177]	other	4	800	human adjudicated	15	some concerns
Qin et al. [176]	other	2	200	administrative field; task outcome; none	6	some concerns
Rafe and Das [179]	document coding	4	2,416	human adjudicated; none; administrative field	24	some concerns
Rafe and Das [180]	extraction	1	590	human adjudicated; none	7	low
Rafe and Das [181]	social-science annotation	3	79,510	human adjudicated; llm teacher; none	13	low
Rahimi et al. [182]	other	1	30	task outcome	6	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Rajput et al. [183]	classification	2	299,463	human single	18	high
Rao and Callison-Burch [184]	judging	7	5,003	human single; human crowd; none	27	high
Ren et al. [186]	safety	8	577	administrative field	14	high
Rezvani et al. [187]	classification; safety	1	1,916	administrative field; task outcome	14	high
Robitza [188]	scientific workflow	2	1,936	none; human crowd	14	some concerns
Sadoun et al. [190]	other; retrieval	2	12,280	task outcome; synthetic; human single; none; human adjudicated	14	some concerns
Saeed et al. [191]	classification	15	500	administrative field	17	high
Sakhawat and Raiyan [192]	classification	3	3,265	human adjudicated; synthetic; none	13	some concerns
Santillana [193]	classification	4	2,021	administrative field; none	13	some concerns
Sautenkov et al. [194]	agent control	1	255	synthetic	7	high
Severo et al. [195]	classification	1	6,000	administrative field	14	high
Shankaranarayana et al. [196]	other; QA; classification; safety	1	1,680	task outcome; human adjudicated; synthetic	24	some concerns
Sharma and Lall [198]	memory	4	778	llm teacher; human single; administrative field	17	high
Sharma et al. [197]	classification	3	500	synthetic	10	high
Sheng et al. [199]	agent control	1	100	task outcome	10	some concerns
Shi et al. [200]	judging	3	579	human single; none	13	high
Song et al. [201]	classification	4	231	not reported; synthetic	14	some concerns
Song et al. [202]	scientific workflow	11	6,466	human adjudicated; human crowd; llm teacher	16	high
Song et al. [203]	QA	5	1,700	human adjudicated	18	high
Sonthalia et al. [204]	classification	1	1,000	not reported	10	some concerns
Stambler [207]	other	6	47	synthetic; administrative field; human single	19	high
Steiner et al. [208]	classification	7	30,225	other: benchmark gold match labels (curated by benchmark authors)	22	some concerns
Su and Huang [209]	agent control	2	100	task outcome	10	some concerns
Sudjianto and Burakov [210]	classification	7	10,000	administrative field	9	some concerns
Sun et al. [211]	QA; retrieval	2	105	human adjudicated; none	10	high
Sun et al. [212]	classification	4	1,200	administrative field	22	some concerns
Tan et al. [213]	safety	1	100	llm teacher	7	high
Tang and Chen [214]	classification; extraction	1	200	human single	14	high
Teke [215]	classification	4	84	model consensus; human single	15	high
Teke [216]	classification	3	4,275	model consensus; none; human single	20	high
Teng et al. [217]	classification	6	2,521	not reported	15	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Terven et al. [218]	retrieval	6	40	llm teacher; other: human-annotated reference evidence (QASPER); other: human-annotated unanswerable (QASPER)	14	some concerns
Tommasino and Rinaldi [220]	classification	5	12,692	human single	16	some concerns
Minh et al. [154]	classification	2	163	human single	12	high
Vats [231]	agent control	1	72	task outcome; none	10	some concerns
Wake and Wagle [233]	agent control	2	60	none; synthetic; task outcome	16	some concerns
Wang and Zang [242]	other	3	200	synthetic	11	high
Wang et al. [236]	classification; other	1	86	administrative field; task outcome	14	high
Wang et al. [241]	extraction	2	1,953	not reported	25	some concerns
Wang et al. [246]	classification	7	207,900	human crowd; none	16	some concerns
Wang et al. [238]	safety	5	100	synthetic; administrative field; llm teacher; vendor label	15	high
Wang et al. [240]	agent control	3	278	human single; none; task outcome	7	some concerns
Wang and Gao [243]	safety	1	5,219	administrative field	12	some concerns
Wang et al. [237]	agent control	1	150	task outcome; not reported; synthetic; none	13	high
Wang et al. [239]	classification	5	19,477	task outcome	9	some concerns
Wang [234]	retrieval; memory	4	1,540	administrative field; llm teacher; none	9	high
Wang et al. [245]	other	3	194	administrative field; not reported	10	some concerns
Wang et al. [244]	QA	3	198	human adjudicated	13	high
Wang et al. [235]	other	9	1,858	administrative field	16	some concerns
multimodalart [157]	multi-benchmark leaderboard; intent classification	5	3,080	human assigned	7	not graded
Weng et al. [247]	classification	8	111	not reported; llm teacher; none	11	high
Wu et al. [252]	agent control	8	100	task outcome	25	some concerns
Wu and Lim [253]	safety	1	510	task outcome	13	high
Wu et al. [251]	safety	1	60	synthetic	13	high
Wu and Lim [254]	agent control	8	100	human adjudicated; task outcome	14	some concerns
Wu [255]	retrieval	2	1,000	administrative field; synthetic; none	9	high
Xie et al. [256]	classification; agent control	7	500	synthetic	21	high
Xu and Mei [258]	classification; other	3	100	administrative field	11	some concerns
Xu et al. [260]	safety	3	2,657	task outcome; human crowd	11	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Xu et al. [257]	judging; QA	1	192	llm teacher; human single; task outcome; none	12	high
Xu et al. [259]	safety	3	3,072	administrative field; synthetic; human single; none	12	some concerns
Yadav [261]	safety	1	142,658	human crowd	14	high
Yamamoto and Kuroki [262]	judging	4	188	human adjudicated; human crowd; task outcome; none	10	high
Yan [263]	classification	4	4,300	administrative field; none	12	high
Yan [264]	classification	1	48	administrative field; human single	7	high
Yang et al. [267]	QA; agent control	1	134	human adjudicated; task outcome	12	some concerns
Yang et al. [266]	other	4	1,120	task outcome; administrative field	16	some concerns
Yang et al. [268]	safety	7	1,440	synthetic	12	high
Yang et al. [265]	other	3	738	synthetic; administrative field	15	some concerns
Ye et al. [269]	other	5	512	synthetic	15	high
Yilmaz et al. [270]	classification	2	400	synthetic; administrative field; human crowd; task outcome; none	17	some concerns
You et al. [271]	classification	1	135	administrative field; human crowd	15	high
Yu and Yao [272]	classification	8	7,532	human adjudicated	10	high
Zhan et al. [273]	classification; agent control	12	231	not reported; task outcome	20	some concerns
Zhang et al. [277]	classification; other	5	213	human crowd; human single; none	14	high
Zhang et al. [280]	other; extraction	1	14,262	human single	10	some concerns
Zhang and Yu [281]	classification	4	2,210	human crowd	15	high
Zhang et al. [274]	classification	2	963,408	administrative field; human single; none; task outcome; llm teacher; model consensus	15	some concerns
Zhang et al. [278]	agent control	1	1,192	task outcome	22	some concerns
Zhang [279]	agent control	1	not reported	task outcome; none	5	high
Zhang et al. [276]	classification	8	2,091	human adjudicated	22	high
Zhang et al. [275]	classification; QA; safety	6	5,059	administrative field; human crowd; none	23	high
Zhao et al. [282]	other: graph anomaly detection	3	10,224	other: benchmark anomaly labels (injected and real); none	7	some concerns
Zheng and Chen [283]	QA	2	768	administrative field	13	high
Zhong et al. [285]	other	1	128	synthetic	12	high
Zhou et al. [286]	classification	8	3,669	human single; human crowd; none	16	some concerns
Zhou et al. [287]	agent control	1	21	task outcome; none	10	high

Study	Task family	Models	n	Reference labels	Rows	RoB
Zhu et al. [289]	classification	5	111	administrative field; synthetic; none	14	high
Zhu et al. [288]	agent control	4	500	synthetic; not reported; task outcome; none	15	high
Zhu et al. [290]	other; QA	2	591	human adjudicated	10	high
Zhuang et al. [291]	other	8	32,233	synthetic; administrative field	15	some concerns
Zhuang et al. [292]	classification	6	7,800	human adjudicated; synthetic; none	15	high

Table S10. The seven application areas of Section 10, the ledger studies that support each one, the kind of reference label each study used and the risk-of-bias grades the area received.

Area	Supporting studies	Reference types	Grade distribution
Document coding at population scale	[179], [276]	human adjudicated; none; administrative field	high (1), some concerns (1)
Annotation for science	[80], [44], [188], [104]	human adjudicated; none; human crowd	high (2), some concerns (2)
Judging and evaluation	[116], [76]	vendor label; none; human adjudicated; task outcome	some concerns (2)
Agent control and memory	[254], [84], [230], [51]	human adjudicated; task outcome; llm teacher; model consensus; none; vendor label	high (3), some concerns (1)
Guardrails, routing and edge orchestration	[101], [102], [186], [65], [185]	task outcome; administrative field; human adjudicated; llm teacher; none	high (2), some concerns (3)
Games and real-time control	[142], [32]	task outcome	high (1), some concerns (1)
Vision and multimodal	[272]	human adjudicated	high (1)

Table S11 sets each launch claim of the vendor, with its own qualifiers, beside its status in the independent records of the ledger under the status rules of Section 7.1.

References

[1] S. M. Mahbub Uddin Ahmed. 2026. Is a 9B Decision Foundation Model Needed for Symptom-to-Disease Classification? Fine-Tuning Clef-Flash Against TF-IDF Baselines. arXiv:2610.15029 [cs.CL] https://arxiv.org/abs/2610.15029

[2] Guy Amit. 2026. Permutation-Robust Decision Modeling with Candidate-Independent Block-Causal Attention. arXiv:2610.01601 [cs.LG] https://arxiv.org/abs/2610.01601

[3] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2022. Conformal Risk Control. arXiv:2208.02814 [stat.ME] https://arxiv.org/abs/2208.02814

[4] Carlos André Antunes, Aldry Albuquerque, Gabriel de Melo Schmith, Natã Batista, Pedro Oziel Escobar da Silva, Stefano Benko Rossi, Lingrid Dietrich Lorenz da Silva, Flávia Cristina da Costa, Thiago Bertotto, and Luis Fernando Fausto. 2026. Narrow Decision Models: A System-1 Architecture for Calibrated, Typed Decisions. arXiv:2610.25553 [cs.AI] https://arxiv.org/abs/2610.25553

[5] Bushra Asseri and Abdulaziz Asseri. 2026. Calibrated to Whom? Persona and Language Effects on Cultural Values in JEV. arXiv:2609.36399 [cs.CL] https://arxiv.org/abs/2609.36399

[6] Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. 2026. Labels Override Definitions in Jev-Style Typed Decision Models. arXiv:2610.02586 [cs.AI] https://arxiv.org/abs/2610.02586

[7] Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. 2026. One Word Opens the Gate: The Option-Channel Attack on Typed Decision Models as Agent Guardrails. arXiv:2610.12292 [cs.AI] https://arxiv.org/abs/2610.12292

Table S11. The vendor's launch claims against their independent status. Squares count the independent records that agree with a claim (■), agree in part (◪) or contradict it (□).

Vendor claim and source	Status and records	Independent evidence
Faster than frontier language models, by 40× to 200× on System One shaped queries in the launch post, and by 193.6× on the homepage, which the vendor expects to lie at the higher end of real-world gains [224, 229]	Supported in direction ■3 ◪39 □1	Faster in every independent study, with every ratio below the launch-post range and a median of 5.3× [44, 101, 102, 116, 276]
Cheaper than frontier language models, by 444.6× on the homepage, under the same higher-end qualifier [224, 229]	Supported in direction ■1 ◪39 □4	Cheaper in all but one independent row, with a median ratio of 37×, while in one study a lower-priced comparator billed less for the same calls [44, 65, 80, 101, 102, 104, 116, 276]
The model never makes type errors, which the vendor presents as a guarantee of the output schema rather than as a measured rate [224]	Supported ■2 ◪0 □0	No type errors or invalid answers were observed in the two studies that count them [80, 212]
The model cannot hallucinate, and the vendor states that its own zero rate is not empirical because schema matching is guaranteed [224]	Definitional ■0 ◪0 □2	Holds for the answer type, while wrong answers still occur at high confidence, with high-confidence accuracy only 0.012 above the base rate on one task [80, 163]
All answers are accompanied with calibrated probabilities and confidence scores, while the vendor's own documentation describes failure modes for this version [224, 226]	Not supported as stated ■2 ◪5 □1	Raw error spans a 77-fold range across studies, below every comparator in some and above most recalibrated comparators in another [14, 65, 77, 80, 132, 163, 179]
Owns the Pareto frontier of quality, cost and time on the vendor's own workflow evaluations, whose reference answers average two frontier models [224]	Supported in direction ■2 ◪3 □0	Not dominated in any independent study with quality and cost on one workload, and less accurate than that workload's comparator in three of them, while a developer's small replica dominates it at 0.60 of its cost [9, 44, 80, 101, 116, 276]

[8] Gian Luca Bailo. 2026. System Switch: When Should a Fast Decision Model Stop and Think? arXiv:2610.09683 [cs.AI] https://arxiv.org/abs/2610.09683

[9] Balian, Rob. 2026. *rev: Reverse Engineering Jev Using Open Models.* Retrieved 2026-09-24 from https://github.com/robbalian/rev

[10] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, Risk-controlling Prediction Sets. *J. ACM* 68, 6 (2021), 1–34. arXiv:2101.02703 doi:10.1145/3478535

[11] Jaime Bellver-Soler, Iván Martín-Fernández, Anmol Guragain, and Luis Fernando D'Haro. 2026. Audio System One Model: Typed Probabilistic Perception of Emotion in Speech. arXiv:2610.22401 [cs.SD] https://arxiv.org/abs/2610.22401

[12] Gaston Besanson. 2026. Probabilistic Sensing, Deterministic Authority: Admitting Model-Produced Observations into Sufficiency-Checked Governance Contracts. arXiv:2610.10978 [cs.SE] https://arxiv.org/abs/2610.10978

[13] Bespoke Labs. 2026. *Bespoke-Nimble-9B Model Card.* Retrieved 2026-09-24 from https://huggingface.co/bespokelabs/Bespoke-Nimble-9B

[14] Bespoke Labs. 2026. *nimble: Local Typed Decisions, Contrastive Data Curation, and Model Evaluation.* Retrieved 2026-09-24 from https://github.com/bespokelabsai/nimble

[15] Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78, 1 (1950), 1–3. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

[16] Pedro Carvalho Brom. 2026. Confidence Is Not a Promise: A Reference Audit of Typed Decision Models. arXiv:2610.36311 [cs.LG] https://arxiv.org/abs/2610.36311

[17] Yves Bru, Kamaldeep Singh Oberoi, and Benoit Gaudou. 2026. Towards Adaptive and Real-Survey-Aligned Simulated Urban Mobility Behaviour of Generative Agents. arXiv:2610.22407 [cs.MA] https://arxiv.org/abs/2610.22407

[18] Felix Buecheler, Gerardo Berbeglia, and Mingyu Guo. 2026. A Cue's Hidden Price: Auditing Seller Influence on AI Shoppers. arXiv:2610.25843 [cs.AI] https://arxiv.org/abs/2610.25843

[19] Syed Usama Khalid Bukhari, Syed Safwan Khalid, and Muhammad Aamir. 2026. System X: An LLM Decision Model That Replies from Its Own KV Cache. arXiv:2610.25456 [cs.AI] https://arxiv.org/abs/2610.25456

[20] Cai. 2026. *ZefanCai/Open-Jev-9B.* Retrieved 2026-09-24 from https://huggingface.co/ZefanCai/Open-Jev-9B

[21] Congfeng Cao, Lipeng Zuo, Konstantinos Papakostas, Qiwei Xu, Songwei Xu, Qianru Meng, Lun Zhou, Zhaochun Ren, Yougang Lyu, and Xiaohui Yan. 2026. SearchJev: A Calibrated System-1 Model for Search Agents. arXiv:2610.05107 [cs.IR] https://arxiv.org/abs/2610.05107

[22] Hao Chen. 2026. COGNIT-Guard: Calibrated Standalone Direct-Decision Guardrails with Heterogeneous CPU-NPU Confidence Cascading under Explicit Latency and False-Positive Constraints. arXiv:2609.33671 [cs.CR] https://arxiv.org/abs/2609.33671

[23] Han Chen and Yingrui Li. 2026. Probability Contracts: Accuracy, Coherence, and Decisions Across LLM Interfaces. arXiv:2609.37470 [cs.LG] https://arxiv.org/abs/2609.37470

[24] Jianbo Chen. 2026. The Company a Question Keeps: Auditing Open Jev-Style Models for Retrieval. arXiv:2610.14631 [cs.IR] https://arxiv.org/abs/2610.14631

[25] Josef Liyanjun Chen. 2026. RoboJev: An Open Vision-Language Judge for Robot Success Detection. arXiv:2610.33718 [cs.RO] https://arxiv.org/abs/2610.33718

[26] Man Chen, Antai Chen, Shuai Xu, Chengjie Gu, Huaping Zhou, Kelei Sun, Tao Wu, and Yongjie Huang. 2026. RSAJEV: Training-Free Abstaining Prototype Memory for Eliminating Silent Hallucination in JEV-Based Remote Sensing Classification. arXiv:2610.15742 [cs.CV] https://arxiv.org/abs/2610.15742

[27] Peixin Chen, Zhigao Chen, Zhiyun Fan, Ying Fang, Zhifu Gao, Qing Gu, Jiajun He, Kangqi Jing, Xiangang Li, Yalin Li, Baiji Liu, Ying Liu, Zhaoci Liu, Yuezhang Peng, Zhendong Peng, Yu Pu, Peiyao Wang, Wupeng Wang, Yuzhong Wu, Yunchong Xiao, Nan Yang, and Renjie Yu. 2026. Qwen-Audio-Jev: A Unified Interface for Audio Decisions. arXiv:2610.21614 [cs.SD] https://arxiv.org/abs/2610.21614

[28] Tianyu Chen, Xinyang Liu, Ruyi An, Zhendong Wang, and Mingyuan Zhou. 2026. System1-Arena: Which Decisions Can Be Left to System 1? arXiv:2610.21244 [cs.CL] https://arxiv.org/abs/2610.21244

[29] Xupeng Chen and Chenqian Le. 2026. Suggested Is Not Superseded: A Pre-Registered Evaluation of a System-One Model as a Memory-Retirement Gate. arXiv:2610.15710 [cs.AI] https://arxiv.org/abs/2610.15710

[30] Zheng Chen, Fei Yu, Haohao Huang, Yang Li, Anlong Chen, and Lei Chen. 2026. SecJev: Bringing Security Expertise to System One Decision Models. arXiv:2610.03073 [cs.CR] https://arxiv.org/abs/2610.03073

[31] Jianpeng Cheng, Guangyu Sun, Aashu Singh, Benyu Zhang, Haixing Dai, Hossein Mansour, Jiangfan Zhang, Shlok Kumar Mishra, Wei Sun, Xuanming Cui, Yanli Liu, Qi Guo, Max Xiangjun Fan, and Jun Xiao. 2026. MetaEncoder: Exploring the Limit of Bi-Encoders for Multimodal System One Decision Making with Natural Language Interface. arXiv:2610.11316 [cs.CL] https://arxiv.org/abs/2610.11316

[32] Zehua Cheng, Wei Dai, and Jiahao Sun. 2026. this-that-model-1.0: A typed decision model that decides in 30 ms, for a millionth of a cent. arXiv:2609.23886 [cs.CL] https://arxiv.org/abs/2609.23886

[33] Avdhesh Singh Chouhan and Jack Perry. 2026. Verdict: Encode Once, Decide Many. arXiv:2610.20778 [cs.CL] https://arxiv.org/abs/2610.20778

[34] C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16, 1 (1970), 41–46. doi:10.1109/tit.1970.1054406

[35] Kawai Chung, Chunkit Chan, Yuxuan Liu, Yauwai Yim, Xi Yang, Wenbin Hu, Huihao Jing, Chiu Yeung Tang, Changxuan Fan, Ki Sen Hung, Tsun On Kwok, and Haoran Li. 2026. Jev-ToM: Evaluating Theory of Mind in Typed Decision Models. arXiv:2610.18610 [cs.CL] https://arxiv.org/abs/2610.18610

[36] Gabriele Cinà. 2026. Jev Matches 7B Language Models for Speech-Neuroprosthesis Rescoring. arXiv:2609.33538 [cs.CL] https://arxiv.org/abs/2609.33538

[37] Cloudflare. 2026. *Introducing Clef: Our Open-Source Decision Models, and New RL Fine-Tuning Platform.* https://blog.cloudflare.com/clef-decision-models

[38] Convai Innovations. 2026. *convaiinnovations/laya-multilingual.* Retrieved 2026-09-24 from https://huggingface.co/convaiinnovations/laya-multilingual

[39] Convai Innovations. 2026. *Laya Model Card (convaiinnovations/laya).* Retrieved 2026-09-24 from https://huggingface.co/convaiinnovations/laya

[40] Shenghong Dai, Shiva Kumar Pentyala, Yingchi Liu, Shubham Mehrotra, Suman Banerjee, James Zhu, Bin Bi, Sitaram Asur, and Phil Mui. 2026. Koa-decision: Fast and Consistent Structured Decision Making with Generative LLMs. arXiv:2609.36115 [cs.LG] https://arxiv.org/abs/2609.36115

[41] Yuyang Dai, Yuting Tang, Futoshi Futami, and Takashi Ishida. 2026. ProbabilityBench: Model Evaluation and Calibration with Exact Ground-Truth Uncertainty. arXiv:2610.17016 [cs.LG] https://arxiv.org/abs/2610.17016

[42] Pritam Deka. 2026. From Retrieval to Typed Decisions: Calibrated System One Models from Biomedical Sentence Encoders. arXiv:2610.02486 [cs.CL] https://arxiv.org/abs/2610.02486

[43] M. Mikail Demir. 2026. Commit or Abstain? System One Decision Models, LLMs, and Missing Legal Authority. arXiv:2610.29961 [cs.AI] https://arxiv.org/abs/2610.29961

[44] Boyuan Deng, Shuyi Fan, Hongyang Zhang, and Xinhong Xie. 2026. Jev for Scientific Decisions: Evaluating Semantic Choices and Their Consequences. arXiv:2609.24965 [cs.CL] https://arxiv.org/abs/2609.24965

[45] Steven Denney and Matthew DiGiuseppe. 2026. JEV versus LLMs: Accuracy, Cost and Calibration on Seven Political Science Replications. arXiv:2610.06625 [cs.CL] https://arxiv.org/abs/2610.06625

[46] Tobias Deußer, Lorenz Sparrenberg, and Rafet Sifa. 2026. Evaluating and Benchmarking the System One Model Jev. arXiv:2609.37647 [cs.CL] https://arxiv.org/abs/2609.37647

[47] Li Ding, Haidi Jin, and Chen Ji. 2026. Bongard: Training Machine Intuition. arXiv:2609.39111 [cs.CL] https://arxiv.org/abs/2609.39111

[48] Shenghao Ding. 2026. JET: Justification Evaluation in Transformer. arXiv:2609.33874 [cs.LG] https://arxiv.org/abs/2609.33874

[49] Michael Dinzinger, Kanishka Ghosh Dastidar, and Michael Granitzer. 2026. System One models Are Not Better Rerankers. arXiv:2610.32318 [cs.CL] https://arxiv.org/abs/2610.32318

[50] Ben Dong, Tianyi Lu, Zhonghao Guo, and Qian Wang. 2026. DICE: Tilting the Odds through Interface Attacks on System One Model. arXiv:2610.14404 [cs.CR] https://arxiv.org/abs/2610.14404

[51] Joas Antonio dos Santos Barbosa. 2026. Calibrated Decision Models for Autonomous Penetration-Testing Harnesses: JEV and Laya as System One Decision Layers for LLM-Driven Pentest Agents. arXiv:2609.28940 [cs.CR] https://arxiv.org/abs/2609.28940

[52] Yanrui Du, Danyang Zhao, Sendong Zhao, Wenhao Li, Yibo Gao, Haochun Wang, Zhuoer Xu, Shiwen Cui, and Bing Qin. 2026. How Reliable Are System-One Decisions in Medicine? A Comparative Study of Jev, Jev-like Models, and LLMs. arXiv:2610.18529 [cs.CL] https://arxiv.org/abs/2610.18529

[53] Feiyu Duan, Jiayu Lin, Jia Wang, Jun Xiang, Jialiang Wu, Xinnong Zhang, Hanqi Yan, Siyuan Wang, and Zhongyu Wei. 2026. General Decision Models: Benchmarking and Insights Beyond Jev. arXiv:2610.03935 [cs.CL] https://arxiv.org/abs/2610.03935

[54] Charles Elkan. 2001. The foundations of cost-sensitive learning. (2001), 973–978.

[55] Edward S. Epstein. 1969. A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology* 8, 6 (1969), 985–987. doi:10.1175/1520-0450(1969)008<0985:assfpf>2.0.co;2

[56] Romi Fadli, Rusdah, Wahyu Ramadani Putri Utami, and Zulfa Malika Amana. 2026. Does System One Transfer? Calibrated Typed-Decision Models, Label Budgets, and Translation for Indonesian Regional Languages. arXiv:2610.14112 [cs.CL] https://arxiv.org/abs/2610.14112

[57] Yangyang Feng, Jianle Gan, Zhuoyan Feng, and Chengwei Qin. 2026. Jev-DLC: Decision-Level Continual Learning from Typed Feedback. arXiv:2610.31521 [cs.AI] https://arxiv.org/abs/2610.31521

[58] Leandro Carísio Fernandes and Jayr Pereira. 2026. Jev-as-a-Judge for Brazilian Portuguese Information Retrieval: Relevance Assessment and Reranking. arXiv:2610.23346 [cs.IR] https://arxiv.org/abs/2610.23346

[59] FLock.io. 2026. *flock-io/this-that-model-1.0.* Retrieved 2026-09-24 from https://huggingface.co/flock-io/this-that-model-1.0

[60] Shuyu Gan, Young-Jun Lee, and Dongyeop Kang. 2026. SanSi: A Looped Typed Decision Model for System 1.5 Thinking. arXiv:2610.07730 [cs.CL] https://arxiv.org/abs/2610.07730

[61] Tianxiang Gao, Jinzhe Li, Zhiyuan Li, Yi Chang, and Yuan Wu. 2026. More Choices, Fewer Decisions: Ordinal-Scale Bias in JEV-like Direct-Decision Models. arXiv:2609.38827 [cs.AI] https://arxiv.org/abs/2609.38827

[62] Rotem Gershon, Seffi Cohen, Dean Ariel, Havatzelet Bilavsky-Yarden, Amiel Sberro, Shai Yitzhaki, Osnat Tausky, Naama Golan, Maayan Rabinovitch, Ariel Avraham Hasidim, Itay Arianu, Maya Dagan, Ran Balicer, and Noa Dagan. 2026. Decision language models as affordable and rapid alternatives to LLM judges of clinical generative AI. arXiv:2610.28826 [cs.AI] https://arxiv.org/abs/2610.28826

[63] Jonathan Geuter, Andrew Cheng, and David Alvarez-Melis. 2026. Releasing the Inner Jev: A Simple Recipe for Open Decision Models. arXiv:2610.17205 [cs.AI] https://arxiv.org/abs/2610.17205

[64] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378. doi:10.1198/016214506000001437

[65] Ruoqi Guo, Yi Liu, Gelei Deng, Yuekang Li, Lida Zhao, Yutao Wu, Simin Chen, Ying Zhang, and Leo Yu Zhang. 2026. Just Ask Jev: Reinforcement Learning for Calibrated Decisions as a Zero-Shot Detector of AI Alignment Failures. arXiv:2609.29429 [cs.AI] https://arxiv.org/abs/2609.29429

[66] Xiaoguang Guo, Yaohao Long, and Chuxu Zhang. 2026. BLADE: Auditing Information Use for Failure Attribution in LLM Multi-Agent Systems. arXiv:2610.18780 [cs.AI] https://arxiv.org/abs/2610.18780

[67] Phat Huu Ha, Nguyen Xuan Phuc, Nong Nhut Quy, Trinh Tran Tran, Nguyen Minh Triet, and Dang Van Thin. 2026. How Well Do Decision Models Transfer to Vietnamese? An Empirical Study of Text Classification. arXiv:2610.31422 [cs.CL] https://arxiv.org/abs/2610.31422

[68] Jin Han, Mingfeng Xin, Feng Liang, and Jingtao Yu. 2026. NIOCIS: Typed-Decision Next-Hop Routing with Node-Conditioned Option Spaces for Tool-Use Agents. arXiv:2610.25985 [cs.CR] https://arxiv.org/abs/2610.25985

[69] Zhi Han, Zihan Guo, Chenxi Zeng, Zicai Cui, Yue Xin, Ming Zhou, Dongzhan Zhou, Weinan Zhang, and Yang Li. 2026. Embodied Skill Bench: Benchmarking How Well Agents with Embodied Skills Perform on Physical Tasks. arXiv:2610.16238 [cs.RO] https://arxiv.org/abs/2610.16238

[70] Mengxian He, Ziyue Gao, Chun Li, Shunyi Zhang, Hao Wei, Minqing Zhang, and Wu Yuan. 2026. Can JEV Support Clinical Trial Criteria Assessment? arXiv:2610.35721 [cs.LG] https://arxiv.org/abs/2610.35721

[71] Heman10x-NGU. 2026. *Verdict-open-jev: Non-Autoregressive Decision Engine on ModernBERT (151M) with Calibrated Uncertainty (RLCD).* Retrieved 2026-09-24 from https://github.com/Heman10x-NGU/Verdict-open-jev

[72] Jianyi Hu, Hangtao Zhang, Yi Liu, Yeqi Zeng, Li Zeng, Xianlong Wang, Rui Wang, and Leo Yu Zhang. 2026. JevAdvBench: A Benchmark and Black-Box Attacks for Reinforcement Learning for Calibrated Decisions Models. arXiv:2609.31142 [cs.CR] https://arxiv.org/abs/2609.31142

[73] Wenbin Hu, Huihao Jing, Haochen Shi, Yuxuan Liu, Changxuan Fan, Sirui Zhang, Kawai Chung, Hanyu Yang, Zhenhuan Wang, Ziyi Chen, Haoran Li, and Yangqiu Song. 2026. Jev-Guard: Fast Routing and Auditing Agent Actions over Editable Policy Trees. arXiv:2610.18326 [cs.CR] https://arxiv.org/abs/2610.18326

[74] Jerrold Huang, Junwei Quan, Saiyue Lyu, Zihan Guo, and Yingqiang Ge. 2026. SkillScreen: High-Recall Screening of Malicious Agent Skills with Decision Models. arXiv:2610.17727 [cs.CR] https://arxiv.org/abs/2610.17727

[75] Jianguo Huang, Yao Long Teng, Yanchen Deng, and Bo An. 2026. Consider, Then Choose: Order-Invariant Decisions in System One Models. arXiv:2610.18233 [cs.AI] https://arxiv.org/abs/2610.18233

[76] Jiaju Huang, Hao Yang, Xinyu Ma, Xinglong Liang, Kunyan Cai, Junqiang Ma, Shaobin Chen, Yue Sun, and Tao Tan. 2026. Can Jev Judge Radiology Reports? Evaluating a System One Model for Clinical Factuality. arXiv:2609.27607 [cs.CL] https://arxiv.org/abs/2609.27607

[77] Hume, Archer. 2026. *Jev's Architecture Unmasked.* Retrieved 2026-09-24 from https://archerhume.com/posts/jevs-architecture-unmasked/

[78] Wenhua Huo, Fenglei Han, Wangyuan Zhao, Jialin Wu, and Jiayi Han. 2026. NL2Hull: A Natural Language-Driven Constrained Ship Design Decision Framework. arXiv:2610.09896 [cs.AI] https://arxiv.org/abs/2610.09896

[79] iApp Technology. 2026. *iapp/OpenThai-SystemOne.* Retrieved 2026-09-24 from https://huggingface.co/iapp/OpenThai-SystemOne

[80] Hazem Ibrahim and Yasir Zaki. 2026. Evaluating Decision Models for Text Annotation in Computational Social Science. arXiv:2609.24574 [cs.CL] https://arxiv.org/abs/2609.24574

[81] Yuma Ichikawa. 2026. Decision Model Factory: Building Specialized Sovereign Decision Models. arXiv:2610.34956 [cs.AI] https://arxiv.org/abs/2610.34956

[82] Ilaria Imperatrice, Taisia Kameneva, Adam Eric Greenberg, and Alessandro Romano. 2026. Are AI Decision Models Biased? A Case Study of Jev. arXiv:2610.19619 [cs.AI] https://arxiv.org/abs/2610.19619

[83] Abdul Rahman Iqbal, Muhammad Fawad Nasim, and Amir Saleem. 2026. Typed Decision Models under Prevalence Shift: Decision Training, Calibration, and Implicit Priors. arXiv:2610.26325 [cs.LG] https://arxiv.org/abs/2610.26325

[84] Dongming Jiang, Yi Li, and Bingzhe Li. 2026. Jev-Mem: System-One-Controlled Agentic Memory for Efficient AI Agents. arXiv:2609.23986 [cs.AI] https://arxiv.org/abs/2609.23986

[85] Saman Sarker Joy. 2026. Do System One Decisions Add Up? A Study of Probabilistic Coherence. arXiv:2609.33971 [cs.CL] https://arxiv.org/abs/2609.33971

[86] Wooyoung Jung. 2026. Where Can a Decision Model Diagnose HVAC Faults? Reasoning Demand, Physical Representation, and Robustness Under Shift. arXiv:2610.09937 [cs.AI] https://arxiv.org/abs/2610.09937

[87] Parv Khatri, Harshul Jain, Mehak Kapoor, and Tanya Sah. 2026. Do We Need an LLM to Know Why Sources Conflict? Decision Models for Knowledge-Conflict Typing in RAG. arXiv:2610.36915 [cs.CL] https://arxiv.org/abs/2610.36915

[88] Tina Khezresmaeilzadeh, Seyedarmin Azizi, Erfan Baghaei Potraghloo, Mehdi Kamal, and Massoud Pedram. 2026. Claim Laundering: When Agent Decision Models Trust Records Over Reality. arXiv:2610.25947 [cs.AI] https://arxiv.org/abs/2610.25947

[89] Orhan Konak. 2026. More Features Are Not More Evidence: Limits of Training-Free Human Activity Recognition with Jev. arXiv:2609.36154 [cs.AI] https://arxiv.org/abs/2609.36154

[90] kotoba-lang. 2026. *com-kotobalabs/open-jev-deberta-v3-large.* Retrieved 2026-09-24 from https://huggingface.co/com-kotobalabs/open-jev-deberta-v3-large

[91] Michail-Alexandros Kourtis and George Xilouris. 2026. Type-Safe Decision Frameworks for Agentic 5G Control: A Theory-Driven Testbed Characterization of Where They Can Be Applied. arXiv:2609.33689 [cs.NI] https://arxiv.org/abs/2609.33689
[92] Ted Kwartler, Alan Aqrawi, and Arian Abbasi. 2026. Harness Tokenomics: A Router for the Enterprise Agentic Control Plane. arXiv:2609.28919 [cs.AI] https://arxiv.org/abs/2609.28919
[93] Tianyou Lai, Wentao Yue, Zhiying Deng, and Qilei Li. 2026. Correct Choices, Failed Sequences: Choice2Seq Benchmarks Correctness Retention Across Output Interfaces. arXiv:2610.16191 [cs.AI] https://arxiv.org/abs/2610.16191
[94] Tianyou Lai, Wentao Yue, Zhiwei Zuo, and Qilei Li. 2026. Which Differences Matter? Controlled Tests of Rule Transfer, Joint Reliability, and Information Use in Decision Models. arXiv:2610.30045 [cs.AI] https://arxiv.org/abs/2610.30045
[95] Qianlong Lan, Anuj Kaul, Vinothini Pandurangan, and Indranil Sanyal. 2026. From Detection to Decision: Calibration and Deferral for Prompt-Injection Detection. arXiv:2610.20907 [cs.CR] https://arxiv.org/abs/2610.20907
[96] Yifan Lan, Zhongye Liu, Yurui Chang, Hanyu Wang, Yongkang Du, Lu Lin, and Jinghui Chen. 2026. DecisionRobust: Benchmarking Robustness for JEV Style Decision Models Across Text and Vision. arXiv:2610.19118 [cs.CL] https://arxiv.org/abs/2610.19118
[97] Simon A. Lee, Jiahang Sha, Moonseong Jeong, Yunbei Pan, and Jeffrey N. Chiang. 2026. MedJev: Evaluating Zero-Shot System-One Decision Making in Medicine. arXiv:2610.36006 [cs.CL] https://arxiv.org/abs/2610.36006
[98] Unggi Lee and Haeun Park. 2026. Can a System-One LLM Perform Knowledge Tracing When Few or No Learners Are Logged? arXiv:2610.11135 [cs.CL] https://arxiv.org/abs/2610.11135
[99] Baoteng Li, Wenzhuo Wu, Kongming Liang, and Zhanyu Ma. 2026. Visual Jev Rewards: Reference-Bound Verification for Multi-Subject Image Generation. arXiv:2610.09328 [cs.CV] https://arxiv.org/abs/2610.09328
[100] Delong Li, Chen Li, Xu Wang, Haochen Gong, Rui Lang, and Guangsheng Yu. 2026. SoK: Semantic Decision Engines in Network Control Loops. arXiv:2610.06425 [cs.NI] https://arxiv.org/abs/2610.06425
[101] Delong Li, Xu Wang, Haochen Gong, Rui Lang, and Guangsheng Yu. 2026. Intent Interpretation at RIC Timescales: Jev Decision Models versus Large Language Models in 6G Open RAN. arXiv:2609.23136 [cs.NI] https://arxiv.org/abs/2609.23136
[102] Delong Li, Xu Wang, Haochen Gong, Rui Lang, and Guangsheng Yu. 2026. Replacing Large Language Models with Jev Decision Models for Low-Latency Edge Service Orchestration. arXiv:2609.22753 [cs.DC] https://arxiv.org/abs/2609.22753
[103] Fengxin Li. 2026. Seeing the Options: Typed Decisions over Visual Candidates with Reusable Option Memory. arXiv:2610.33702 [cs.AI] https://arxiv.org/abs/2610.33702
[104] Hengyu Li. 2026. KITE: Scaling Jev Population Experiments with Sparse Flagship Calibration. arXiv:2609.27535 [cs.MA] https://arxiv.org/abs/2609.27535
[105] Jiawei Li. 2026. Fast Models, Slow Evidence: A Paired and Self-Audited Evaluation of System-1 Decision Models for LLM Agent Harnesses. arXiv:2610.02267 [cs.AI] https://arxiv.org/abs/2610.02267
[106] Keyi Li, Yihao He, and Quanyi Li. 2026. Beyond Calibration: Do a Typed-Decision Model's Probabilities Obey the Probability Axioms? arXiv:2609.33209 [cs.CL] https://arxiv.org/abs/2609.33209
[107] Qi Li and Xinchao Wang. 2026. Jeva: System One Models from First Principles. arXiv:2610.18773 [cs.AI] https://arxiv.org/abs/2610.18773
[108] Qiaomu Li and Ying Xie. 2026. Can System-1-Style Decision Models Replace Dedicated Rerankers in RAG? A Systems Evaluation. arXiv:2610.34816 [cs.LG] https://arxiv.org/abs/2610.34816
[109] Qi Li, Jianjun Xu, and Pingtao Wei. 2026. DKnown-Jev: Fine-Tuning and Evaluation of a JEV-Style Model for Tasks Involving Normative Knowledge. arXiv:2610.24358 [cs.AI] https://arxiv.org/abs/2610.24358
[110] Ran Li and Lei Chen. 2026. JevCascade: Readout Stability and Inference-Time Compute Allocation for Prefill-Only Decision Models. arXiv:2610.07716 [cs.CL] https://arxiv.org/abs/2610.07716
[111] Xuechen Li. 2026. AgriJev: Efficient Agricultural Candidate Scoring with Multi-Scale Adapted Models. arXiv:2610.19478 [cs.CL] https://arxiv.org/abs/2610.19478
[112] Xing Li, Qingcheng Chang, Jinzhong Ning, Changfeng Xu, Shenlong Zhang, Yijia Zhang, Ling Luo, and Hongfei Lin. 2026. Specialized Decision Models vs. General-Purpose LLMs: Benchmarking Jev Across Knowledge, Reasoning, and Multilingual Tasks. arXiv:2610.11978 [cs.CL] https://arxiv.org/abs/2610.11978
[113] Xiang Li, Aichun Huang, Mingming Zhang, and Zheli Liu. 2026. Same Attack, Different Judge: A First Controlled Adversarial Comparison of Jev, LLM Judges, and Purpose-Built Safety Classifiers. arXiv:2610.17820 [cs.CR] https://arxiv.org/abs/2610.17820
[114] Xing Li, Jinzhong Ning, Yijia Zhang, Liang Yang, and Hongfei Lin. 2026. Can Decision Models Understand Stance? Evaluating Jev Against General-Purpose LLMs. arXiv:2610.11901 [cs.CL] https://arxiv.org/abs/2610.11901
[115] Yubo Li, Ramayya Krishnan, and Rema Padman. 2026. Reading Right, Acting Wrong: Diagnosing Decision Failures of System One Models in Agent Loops. arXiv:2610.18579 [cs.AI] https://arxiv.org/abs/2610.18579

[116] Yubo Li, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2026. JEV-as-a-Judge: Accept When Confident, Escalate When Unsure. arXiv:2609.26550 [cs.AI] https://arxiv.org/abs/2609.26550
[117] Yinheng Li and Justin Wagle. 2026. LLM-as-Jev: LLMs Are Already Jev-Style Decision Models – When and How to Fine-Tune Them. arXiv:2610.02076 [cs.CL] https://arxiv.org/abs/2610.02076
[118] Yang Li, Yike Wang, Di Wan, Chujun He, Zongsi Xu, Yitao Yu, Quanfeng Lu, Xiyu Chen, Sheng Zhong, Deng Binjie, Yichen Guo, Zhenwei Guo, Hou Jun, Kunhan Guo, Weihao Lin, Tuo Zhou, Hai Liu, Chenwei Wu, Jianyuan Zhong, Xilin Dai, and Ye Luo. 2026. S^2-MAS: A Society-Scale Multi-Agent Simulation Framework with 500 Million Household Persona Agents. arXiv:2610.27746 [cs.MA] https://arxiv.org/abs/2610.27746
[119] Yida Lin. 2026. PACT: Pairwise-Anchored Calibrated Tuning for Single-Token Typed Decisions. arXiv:2609.35865 [cs.CL] https://arxiv.org/abs/2609.35865
[120] Guoming Ling, Muen Xue, and Zijian Ye. 2026. Jev in the Wild: A Data-Driven Analysis of the Jev Model's Functionality, Applications and Ecosystem. arXiv:2609.30216 [cs.SE] https://arxiv.org/abs/2609.30216
[121] Yuchen Ling, Shengcheng Yu, Zhenyu Chen, and Chunrong Fang. 2026. When Faster Decisions Do Not Make Better Agents: An Empirical Study of Jev in Agent Workflows. arXiv:2610.13092 [cs.AI] https://arxiv.org/abs/2610.13092
[122] Bo Liu, Lianyu Wang, Meng Wang, Ching-Yu Cheng, and Huazhu Fu. 2026. RetiJev: Fast and Source-Grounded Ophthalmic Decision Models. arXiv:2610.30947 [cs.CL] https://arxiv.org/abs/2610.30947
[123] Chenghao Liu, Wei Fan, Jindong Han, Dongjie Wang, Aniket Mahanti, Pengyang Wang, and Hao Liu. 2026. TimesJev: Thinking Fast on Time Series with System 1 Sufficiency. arXiv:2610.26002 [cs.AI] https://arxiv.org/abs/2610.26002
[124] Daoxuan Liu, Jiayi Liao, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. 2026. Decision-Native Recommendation with Large Language Models. arXiv:2610.34124 [cs.IR] https://arxiv.org/abs/2610.34124
[125] Fangcheng Liu, Yeqing Shen, Anda Cheng, Weishi Mi, Chao Tang, Chenyuan Liu, Yushun Xiang, Tingguang Li, Yong-Lu Li, and Yehui Tang. 2026. RoboICL: Embodied In-Context Learning with GPT-6 Astra. arXiv:2609.34261 [cs.RO] https://arxiv.org/abs/2609.34261
[126] Fangyuan Liu, Sirui Zhao, Xiaohao Wang, Guo Cheng, Peng Zhang, and Enhong Chen. 2026. PsyJev: Psychological Analysis via System-1 Judgment. arXiv:2610.36689 [cs.AI] https://arxiv.org/abs/2610.36689
[127] Hongliang Liu. 2026. Adversarial Cues in Decision Models Used as Judges: The Role of Request Presentation. arXiv:2610.11436 [cs.CL] https://arxiv.org/abs/2610.11436
[128] Haozheng Liu, Zhaoqian Xue, Zhongyin Wang, Junting Mu, Zihui Ma, Shuya Feng, and Lingyao Li. 2026. Decision Models as Judges for Hallucination Detection. arXiv:2610.37229 [cs.CL] https://arxiv.org/abs/2610.37229
[129] Yixuan Liu. 2026. Evaluating System One Models for Agent Security Decisions: Reliability, Calibration, and Selective Automation. arXiv:2609.33401 [cs.CR] https://arxiv.org/abs/2609.33401
[130] Zhuchenyang Liu, Yao Zhang, and Yu Xiao. 2026. Multimodal Reranking Is a System One Decision. arXiv:2610.28435 [cs.IR] https://arxiv.org/abs/2610.28435
[131] llm-semantic-router. 2026. *llm-semantic-router/Decision-1.0-Sol-2B.* Retrieved 2026-09-24 from https://huggingface.co/llm-semantic-router/Decision-1.0-Sol-2B
[132] LocalLLaMA. 2026. *LocalLLaMA/typed-decisions.* Retrieved 2026-09-24 from https://huggingface.co/datasets/LocalLLaMA/typed-decisions
[133] Michele Loi. 2026. The Argument and the Letterhead: Source-Position Coherence in AI Evaluation. arXiv:2609.35286 [cs.AI] https://arxiv.org/abs/2609.35286
[134] Jiawen Lu and Tongtong Wu. 2026. Benchmarking candidate coverage and rejection policy transfer in typed decision models. arXiv:2610.03387 [cs.AI] https://arxiv.org/abs/2610.03387
[135] Federico Lucca. 2026. Measure Before You Train: Typed Decision Models vs. a Classical Baseline vs. a Generative LLM for Routing a Personal Multilingual Inbox. arXiv:2610.32955 [cs.CL] https://arxiv.org/abs/2610.32955
[136] Fabian Lukassen, Christoph Weisser, and Alexander Silbersdorff. 2026. Cheap, Consistent, Calibrated, but Biased? Typed Decision Models as Judges. arXiv:2610.18122 [cs.CL] https://arxiv.org/abs/2610.18122
[137] Mantas Lukauskas and Viktorija Šarkauskaitė. 2026. Probable but Not Valid: A Psychometric Audit of Decision Models as Synthetic Survey Respondents. arXiv:2610.14912 [cs.AI] https://arxiv.org/abs/2610.14912
[138] Yang Luo, Tiantian Ji, Zifeng Kang, Yucheng Wang, Yong Liu, Zhiyi Yin, and Xinran Liu. 2026. Nudge the Judge: Uncovering Program Decision Hijacking in Real-World Jev Applications. arXiv:2610.18612 [cs.CR] https://arxiv.org/abs/2610.18612
[139] Hanjia Lyu and Yinglong Xia. 2026. Decision-Oriented Recommendation Reranking: An Empirical Study of Its Promise and Limitations. arXiv:2609.40241 [cs.IR] https://arxiv.org/abs/2609.40241
[140] Jie Ma, Jie Gao, Yihang Liu, Zhike Qiu, Junle Li, Chongyi Zhuang, Jiayi Ji, and Xiaoshuai Sun. 2026. FastJEV: Understanding Redundancy for Compact JEV Inference. arXiv:2610.11379 [cs.CV] https://arxiv.org/abs/2610.11379
[141] Linhai Ma, Yan Wang, Lingfei Qian, Yupeng Cao, and Victor Gutierrez Basulto. 2026. Judges Follow the Reference: The Derivation Gap in Verifying Tool-Call Arguments. arXiv:2610.18497 [cs.AI] https://arxiv.org/abs/2610.18497

[142] Weiyu Ma, Liangbing Zhao, Yongcheng Zeng, and Jian Zhao. 2026. JEV-Star: Fast, Low-Cost StarCraft II Control with Language-Model Planning. arXiv:2609.27331 [cs.GT] https://arxiv.org/abs/2609.27331
[143] Xiaoxiao Ma, Weiyang Jin, Yanwen Ma, Haoran Li, Zhixiang Wei, Songchun Zhang, Yaowei Li, Junhao Zhuang, Zeyue Xue, Yuming Li, Zheming Fu, Haoyang Huang, Nan Duan, and Feng Zhao. 2026. WorldStream: Streaming Interactive 3D Worlds with Hierarchical Planning and Persistent State. arXiv:2610.28827 [cs.CV] https://arxiv.org/abs/2610.28827
[144] Yi Ma, Tianpei Yang, Yaodong Yang, Weixun Wang, and Hongyao Tang. 2026. Can Jev be Your Q or Policy in Reinforcement Learning? arXiv:2610.11692 [cs.LG] https://arxiv.org/abs/2610.11692
[145] MachineFi. 2026. *Trio-Spark Model Card.* https://huggingface.co/MachineFi/Trio-Spark
[146] Alfredo Madrid-García and Beatriz Merino-Barbancho. 2026. Jev in Medicine: A Benchmark Evaluation. arXiv:2609.34024 [cs.AI] https://arxiv.org/abs/2609.34024
[147] Mapika. 2026. *Mapika/decider-2b.* Retrieved 2026-09-24 from https://huggingface.co/Mapika/decider-2b
[148] Federico Mazzoni and Andrea Failla. 2026. Benchmarking System One Models in Online Moderation. arXiv:2610.07953 [cs.SI] https://arxiv.org/abs/2610.07953
[149] Haocheng Mei, Yuchuan Tian, Ye Yuan, Enhua Wu, Chao Xu, Allen Chen, Kai Han, and Yunhe Wang. 2026. Improving JEV Entailment Reranking with Span Readouts and Selective Adaptation. arXiv:2610.35654 [cs.CL] https://arxiv.org/abs/2610.35654
[150] Soundouss Messoudi and Youness Mansar. 2026. Calibration and Risk-Budgeted Decisions with System One Models. arXiv:2610.19811 [stat.ML] https://arxiv.org/abs/2610.19811
[151] Microsoft. 2026. *Deploy and Use Microsoft-Decision-1 in Microsoft Foundry.* https://learn.microsoft.com/en-us/azure/foundry/foundry-models/how-to/use-foundry-models-microsoft-decision
[152] Microsoft. 2026. *Introducing Microsoft-Decision-1.* https://commandline.microsoft.com/microsoft-decision-1-model-foundry/
[153] Guanghui Min and Chen Chen. 2026. Prune the Summary, Not the History: Source-Judged Compaction for Long-Horizon Agents. arXiv:2610.20927 [cs.CL] https://arxiv.org/abs/2610.20927
[154] Tran Cao Minh, Hai Tran, and Ha Minh Tan. 2026. Can Multimodal Decision Models Classify Breast Ultrasound Images? Benchmarking CLEF and CLEF-Flash Across Prompts and Datasets. arXiv:2610.13675 [eess.IV] https://arxiv.org/abs/2610.13675
[155] Mohamad Mohebifar and Alex Bit. 2026. OneCodeReviewBench: Typed Decisions for Code Review and Search. arXiv:2610.33389 [cs.SE] https://arxiv.org/abs/2610.33389
[156] Tianwei Mu, Shengyan Jiang, Mingzhe Yuan, Qing Luo, Min Xiao, Wenhong Wang, Jun Li, and Manhong Huang. 2026. HydroJEV: A one-second, training-free screen for cyber-attack and fault attribution in water distribution networks. arXiv:2610.02048 [cs.AI] https://arxiv.org/abs/2610.02048
[157] multimodalart. 2026. *Jev Decision Index.* https://huggingface.co/spaces/multimodalart/jev-decision-index
[158] Rushab Munot. 2026. When Multiple Choices Are Correct: Asymmetric Probability Allocation in Jev and Other Decision Models. arXiv:2610.29217 [cs.AI] https://arxiv.org/abs/2610.29217
[159] Gowthamkumar Nandakishore. 2026. CLM-as-a-Judge: Evaluating an Open Contrastive Decision Model on Public Judge Benchmarks. arXiv:2610.07177 [cs.LG] https://arxiv.org/abs/2610.07177
[160] Gowthamkumar Nandakishore. 2026. Laya as a Typed Probabilistic Assessor: An Independent Reproduction and a Preregistered Study of Calibration and Selective Escalation. arXiv:2609.33843 [cs.AI] https://arxiv.org/abs/2609.33843
[161] Jared Palmer. 2026. *kev: Jev-Like Family of Decision Models Built on Qwen3.5/3.8.* Retrieved 2026-09-24 from https://github.com/jaredpalmer/kev
[162] Palmer, Jared. 2026. *jaredpalmer/kev-0.8b.* Retrieved 2026-09-24 from https://huggingface.co/jaredpalmer/kev-0.8b
[163] Palmer, Jared. 2026. *jaredpalmer/kev-9b.* Retrieved 2026-09-24 from https://huggingface.co/jaredpalmer/kev-9b
[164] Ningxin Pan, Hanyu Li, Yeqing Shen, Ning Ding, Yehui Tang, and Wentao Zhang. 2026. The World Won't Wait: Evaluating and Teaching Agents Under a Live Clock. arXiv:2610.20665 [cs.AI] https://arxiv.org/abs/2610.20665
[165] Ujjawal K. Panchal. 2026. FunDM: Cheap & Label-Free Decision Models from Frozen Language Models. arXiv:2610.17175 [cs.LG] https://arxiv.org/abs/2610.17175
[166] Balamurugan Paramasivam and Sudalai Rajkumar. 2026. RubriX: A Fast, Fine-Tunable Non-Autoregressive Rubric Judge. arXiv:2610.34875 [cs.CL] https://arxiv.org/abs/2610.34875
[167] Demetris Paschalides, George Pallis, and Marios D. Dikaiakos. 2026. To Jev or Not? Evaluating the Accuracy and Efficiency of Structured Decision Models for Hate-Speech Moderation. arXiv:2610.03324 [cs.CL] https://arxiv.org/abs/2610.03324
[168] Chengxuan Pei, Ziyao Wang, Teng Mei, and Zhe Li. 2026. Learning to Formalize: Bridging Open-Ended Workloads and Fast Decision Models. arXiv:2610.36595 [cs.AI] https://arxiv.org/abs/2610.36595
[169] Ismail Pelaseyed and Alan Zabihi. 2026. DFIP: Decision-Focused Inference Path for System One Models. arXiv:2610.36804 [cs.PF] https://arxiv.org/abs/2610.36804

[170] Vasilis Perifanis, Nikolaos Pavlidis, and Symeon Symeonidis. 2026. SeLMRoute: Probabilistic Semantic Evidence for Large Language Model Routing. arXiv:2609.34736 [cs.AI] https://arxiv.org/abs/2609.34736
[171] Perplexity. 2026. *pplx-decider-v1.1-27b Model Card.* https://huggingface.co/perplexity-ai/pplx-decider-v1.1-27b
[172] pngwn. 2026. *pngwn/system-one-qwen3.5-4b-scorer.* Retrieved 2026-09-24 from https://huggingface.co/pngwn/system-one-qwen3.5-4b-scorer
[173] Marco Polignano, Het Darshan Mehta, Pasquale Lops, Ernesto William De Luca, and Giovanni Semeraro. 2026. Decision4Rec: A Jev-Inspired Framework of Decision Models on Knowledge and Active Graphs for Multi-Agent Recommender Systems. arXiv:2610.26035 [cs.AI] https://arxiv.org/abs/2610.26035
[174] Riccardo Porcedda. 2026. Jev thinks "I don't know", but doesn't say it: Introducing Sys1Cal-v1 Dataset for Probability Calibration. arXiv:2609.35342 [cs.AI] https://arxiv.org/abs/2609.35342
[175] Ruizhi Qian, Lei Shi, Tianyu Hu, Bo Yan, Song Wang, and Yu Tian. 2026. SOMA: Continuous Monitoring and Intervention for System Two Reasoning with System One Models. arXiv:2610.26793 [cs.AI] https://arxiv.org/abs/2610.26793
[176] Luozheng Qin, Yuhan Zhu, Changhai Zhou, Qian Qiao, Jun Gao, Yuhua Zhou, Harry Zhang, Xiangyu Zeng, Yuxuan Liu, Yuxin Lu, Kaijie Chen, Andrew Chen, and Pony Ma. 2026. MJev-Spatial: System One Spatial Intelligence via Native Attention Matching. arXiv:2610.22276 [cs.CV] https://arxiv.org/abs/2610.22276
[177] Xukui Qin, Youting Wang, Xinjie He, Ziyang Luo, Runxiong Wu, Yan-Syuan Chen, and Zhongyao Chu. 2026. Decision Readouts for Text-Mediated Video Anomaly Detection: An Exploratory Evaluation of Jev and Qwen. arXiv:2609.34180 [cs.AI] https://arxiv.org/abs/2609.34180
[178] Amir Rafe and Subasish Das. 2026. Benchmarking System One Decision Models against Trained Classifiers and Language Models for Automated Decision Gates. arXiv:2610.00346 [cs.LG] https://arxiv.org/abs/2610.00346
[179] Amir Rafe and Subasish Das. 2026. Calibrated Decisions at Scale: Converting Police Crash Narratives into Probabilistic Crash Variables with a System One Model (Jev). arXiv:2609.24052 [cs.CL] https://arxiv.org/abs/2609.24052
[180] Amir Rafe and Subasish Das. 2026. Counting the Uncounted: Population-Level Surveillance of Documented Pregnancy and Fetal Harm in Police Crash Narratives with a System One Model (Jev). arXiv:2610.00213 [stat.AP] https://arxiv.org/abs/2610.00213
[181] Amir Rafe and Subasish Das. 2026. Estimating Uncoded Crash Factors with Tabular Foundation and System One Models: Kumo Tabular and Jev. arXiv:2610.10321 [stat.AP] https://arxiv.org/abs/2610.10321
[182] Masoud Rahimi, S. M. Matin Alemohammad, Hamid Behroozi, and Mahdi Nouri. 2026. System One Models for Wireless Decision-Making:Applications and Performance Evaluation. arXiv:2610.04345 [eess.SP] https://arxiv.org/abs/2610.04345
[183] Dushyant Rajput, Nirdesh Chauhan, and Siddharth Kosaraju. 2026. Does a Hosted Decision Model Respect Irrelevant Alternatives? A Pre-Registered Test of Option-Set Dependence in Jev. arXiv:2610.16655 [cs.AI] https://arxiv.org/abs/2610.16655
[184] Delip Rao and Chris Callison-Burch. 2026. Decision Models vs. LLMs as Rubric Judges: Cheaper, Faster, and Wrong in the Same Places. arXiv:2609.29769 [cs.CL] https://arxiv.org/abs/2609.29769
[185] Rauch. 2026. *post on X about Jev latency in Vercel AI Gateway.* Retrieved 2026-09-24 from https://x.com/rauchg/status/2100307962262872105
[186] Simiao Ren, Kidus Zewde, Xingyu Shen, Yuchen Zhou, Dennis Ng, Ankit Raj, Tommy Duong, Yuxin Zhang, and Neo Tiangratanakul. 2026. Open-Jev Judgments on CallScreenBench: Calibrated One-Pass Scam Screening with a Small Language Model. arXiv:2609.23959 [cs.CL] https://arxiv.org/abs/2609.23959
[187] Arshak Rezvani, Sasha Behrouzi, and Ahmad-Reza Sadeghi. 2026. JevVibe: Efficient Classification-Guided Secure Code Generation. arXiv:2609.34963 [cs.CR] https://arxiv.org/abs/2609.34963
[188] Werner Robitza. 2026. JEVQA - Video Quality from Metadata, Bitstream, and Pixel Features with a General-Purpose Decision Model. arXiv:2609.24395 [eess.IV] https://arxiv.org/abs/2609.24395
[189] Alden Do Rosario and Kirollos Atef. 2026. Jev-as-a-Judge for RAG Claim Verification: Confidence-Gated Cascading Between a Non-Generative Verifier and a Frontier LLM. arXiv:2610.20634 [cs.CL] https://arxiv.org/abs/2610.20634
[190] Mohamad Yazan Sadoun, Sarah Sharif, and Yaser Mike Banad. 2026. From Probabilities to Decisions: Search and Multi-Teacher Distillation with Jev. arXiv:2610.09188 [cs.AI] https://arxiv.org/abs/2610.09188
[191] Muhammad Atif Saeed, Azka Atiq, Adarsh Kumar, Akhtar Jamil, Uzair Iqbal, and Roland Aydin. 2026. SHIFT-JEV: Certified Abstention and Gain-Gated Continual Adaptation for LLM-Based Decision Models under Distribution Shift. arXiv:2610.34329 [cs.LG] https://arxiv.org/abs/2610.34329
[192] Adib Sakhawat and Syed Rifat Raiyan. 2026. Literal Lock-In: How Test Language Shifts Literal Attraction in Typed Decision Models on Bengali Idioms. arXiv:2610.22526 [cs.CL] https://arxiv.org/abs/2610.22526
[193] Juan S. Santillana. 2026. VX-One: Typed, Calibrated Decision Heads for Real-Time Perimeter Triage. arXiv:2610.30090 [cs.CR] https://arxiv.org/abs/2610.30090

[194] Oleg Sautenkov, Yasheerah Yaqoot, Muhammad Ahsan Mustafa, Grigorii Kalinin, Chih-Yung Wen, and Dzmitry Tsetserukou. 2026. UAV-SwarmAgent: Scalable Multi-UAV Reconnaissance and Task Allocation via Code Generation and Vision-Language Reasoning. arXiv:2610.32559 [cs.RO] https://arxiv.org/abs/2610.32559

[195] Paulo Severo, Silvio E. Quincozes, and Amanda Dias. 2026. Jev-IDS: System One Models for Network Intrusion Detection. arXiv:2610.01079 [cs.CR] https://arxiv.org/abs/2610.01079

[196] Sharath M Shankaranarayana, Davor Runje, and Jan Jannink. 2026. Beyond Answer Confidence: A Controlled Audit of Self-Knowledge in a Black-Box Decision Model. arXiv:2610.01006 [cs.AI] https://arxiv.org/abs/2610.01006

[197] Rahul Sharma, Andrew B. Duncan, Gaétan Marceau Caron, and Sebastian J. Vollmer. 2026. TypedBench: A Benchmark for Calibration, Framing Sensitivity, and Cost in System One Decision Models. arXiv:2610.11392 [cs.AI] https://arxiv.org/abs/2610.11392

[198] Rishabh Sharma and Rishika Lall. 2026. When Does Selection Replace Extraction? A Pre-Registered Test of Agent Memory with a Typed Decision Model. arXiv:2609.34227 [cs.AI] https://arxiv.org/abs/2609.34227

[199] Kai Sheng, Liuyi Wang, Jinlong Li, Haojie Dai, Chengju Liu, and Qijun Chen. 2026. NavJev: Efficient Vision-Language Navigation via Action-Centric Visual Compression and Discriminative Action-Semantic Memory. arXiv:2609.34969 [cs.RO] https://arxiv.org/abs/2609.34969

[200] Xianglong Shi, Shifeng Liu, Sirui Zhao, Shengming Yuan, and Enhong Chen. 2026. Multi-Dimensional Comparative Scale Construction for Efficient Personalized Subjective Judgment in High-Traffic Applications. arXiv:2609.33282 [cs.AI] https://arxiv.org/abs/2609.33282

[201] Dingjie Song, Zhiling Yan, Yuxuan Zhang, and Lichao Sun. 2026. LEV: Thinking Fast and Deep with Looped Decision Models. arXiv:2610.20951 [cs.AI] https://arxiv.org/abs/2610.20951

[202] Dingjie Song, Zhiling Yan, Ziqian Zhang, Yuxuan Zhang, and Lichao Sun. 2026. PremiseCheck: Efficient Fact-Checking of Human and AI Peer Reviews. arXiv:2610.21046 [cs.CL] https://arxiv.org/abs/2610.21046

[203] Mingyang Song, Mao Zheng, Zhiyuan Ma, and Xinpeng Zhou. 2026. Understanding Jev-Style System One Models: Does Structure Reduce Selection Bias? arXiv:2610.35826 [cs.CL] https://arxiv.org/abs/2610.35826

[204] Ankit Sonthalia, Haritz Puerto, Alexander Rubinstein, Martin Gubri, and Seong Joon Oh. 2026. Agent in a Bottle: Can LLM Agents Turn Their Capabilities Into Cheap, Scalable Artifacts? arXiv:2610.08775 [cs.AI] https://arxiv.org/abs/2610.08775

[205] Viknesh Sounderajah, Hutan Ashrafian, Robert M Golub, Shravya Shetty, Jeffrey De Fauw, Lotty Hooft, Karel Moons, Gary Collins, David Moher, Patrick M Bossuyt, Ara Darzi, Alan Karthikesalingam, Alastair K Denniston, Bilal Akhter Mateen, Daniel Ting, Darren Treanor, Dominic King, Felix Greaves, Jonathan Godwin, Jonathan Pearson-Stuttard, Leanne Harling, Matthew McInnes, Nader Rifai, Nenad Tomasev, Pasha Normahani, Penny Whiting, Ravi Aggarwal, Sebastian Vollmer, Sheraz R Markar, Trishan Panch, and Xiaoxuan Liu. 2021. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 11, 6 (2021), e047709. doi:10.1136/bmjopen-2020-047709

[206] Viknesh Sounderajah, Hutan Ashrafian, Sherri Rose, Nigam H. Shah, Marzyeh Ghassemi, Robert Golub, Charles E. Kahn, Andre Esteva, Alan Karthikesalingam, Bilal Mateen, Dale Webster, Dan Milea, Daniel Ting, Darren Treanor, Dominic Cushnan, Dominic King, Duncan McPherson, Ben Glocker, Felix Greaves, Leanne Harling, Johan Ordish, Jérémie F. Cohen, Jon Deeks, Mariska Leeflang, Matthew Diamond, Matthew D. F. McInnes, Melissa McCradden, Michael D. Abràmoff, Pasha Normahani, Sheraz R. Markar, Stephanie Chang, Xiaoxuan Liu, Susan Mallett, Shravya Shetty, Alastair Denniston, Gary S. Collins, David Moher, Penny Whiting, Patrick M. Bossuyt, and Ara Darzi. 2021. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nature Medicine* 27, 10 (2021), 1663–1665. doi:10.1038/s41591-021-01517-0

[207] Lev Stambler. 2026. P(Yes): 15× Improvements via Test-Time Scaling of Decision Models. arXiv:2610.25474 [cs.AI] https://arxiv.org/abs/2610.25474

[208] Aaron Steiner, Ralph Peeters, and Christian Bizer. 2026. Entity Matching using Decision Models. arXiv:2610.34626 [cs.CL] https://arxiv.org/abs/2610.34626

[209] Haoyang Su and Weiran Huang. 2026. JevSpawn: Adaptive Agentic Inference through Compositional Action Spaces. arXiv:2610.00437 [cs.AI] https://arxiv.org/abs/2610.00437

[210] Agus Sudjianto and Denis Burakov. 2026. Teacher Supervision, Fixed Representations and Calibration: Exact Risk Calculations for Candidate-Scoring Decision Models. arXiv:2610.26797 [stat.ML] https://arxiv.org/abs/2610.26797

[211] Yasheng Sun, Yanning Dai, Yimeng Chen, Nanbo Li, Firas Laakom, Yifan Xu, Piotr Piękos, Mateusz Ostaszewski, Wenyi Wang, Cheng Luo, Yuhui Wang, and Jürgen Schmidhuber. 2026. JevGPT: Thinking Fast and Slow in Autonomous Research. arXiv:2610.18272 [cs.CL] https://arxiv.org/abs/2610.18272

[212] Yu Sun, Junhao Xu, Jiajia Shi, and Zijin Yang. 2026. Type-Safe Is Not Error-Free: A Constrained Decision Head Follows the Option Name, Not the Rubric Bound to It. arXiv:2609.26758 [cs.AI] https://arxiv.org/abs/2609.26758

[213] Yixin Tan, Chia-Yi Hsu, Jiayang Liu, Chia-Mu Yu, and Rui Wen. 2026. Diagnose Before You Refine: Multi-Dimensional Judge Feedback for Query-Efficient Jailbreak Attacks. arXiv:2610.19971 [cs.CR] https://arxiv.org/abs/2610.19971

[214] Luyao Tang and Cheng Chen. 2026. OmniMed-Jev: Calibrating LVLM Confidence for Trustworthy Medical Multimodal Decisions via System One. arXiv:2610.00381 [cs.LG] https://arxiv.org/abs/2610.00381
[215] Sait Furkan Teke. 2026. HakemBench: A Turkish Benchmark of Typed Decisions. arXiv:2610.02293 [cs.CL] https://arxiv.org/abs/2610.02293
[216] Sait Furkan Teke. 2026. ufakzeka-karar: An Open Turkish Typed-Decision Model with Order-Invariant Option Scoring. arXiv:2610.06744 [cs.CL] https://arxiv.org/abs/2610.06744
[217] Yao Long Teng, Jianguo Huang, Jiayi Cai, Yanchen Deng, and Bo An. 2026. Prism-1: Position- and Rival-Independent Scoring for System One Models. arXiv:2610.22066 [cs.AI] https://arxiv.org/abs/2610.22066
[218] Juan Terven, Diana-Margarita Córdova-Esparza, Julio-Alejandro Romero-González, Edgar Chavez-Urbiola, Juan Irving Vasquez, Julian Andres Ramirez-Bautista, Silvia L. Chaparro-Cárdenas, and Gendry Alfonso-Francia. 2026. Scan or Index? large Evidence Coverage, Answer Quality, and Cost in Long-Document RAG. arXiv:2610.28957 [cs.CL] https://arxiv.org/abs/2610.28957
[219] TheoLeeCJ. 2026. *SemIf-OpenJev: Semantic Ifs from Open Models.* Retrieved 2026-09-24 from https://github.com/TheoLeeCJ/SemIf-OpenJev
[220] Cristian Tommasino and Antonio M. Rinaldi. 2026. PathJev: Decision-Native Vision-Language Modelling for Histopathology. arXiv:2610.32478 [cs.CV] https://arxiv.org/abs/2610.32478
[221] TypeSafe AI. 2026. *AI Primer.* Retrieved 2026-09-24 from https://docs.typesafe.ai/introduction/machine-learning-primer
[222] TypeSafe AI. 2026. *Box.* https://docs.typesafe.ai/primitives/box.md
[223] TypeSafe AI. 2026. *Confidence.* Retrieved 2026-09-24 from https://docs.typesafe.ai/confidence
[224] TypeSafe AI. 2026. *Introducing System One Models & Jev.* Retrieved 2026-09-24 from https://typesafe.ai/blog/introducing-system-one-models-and-jev
[225] TypeSafe AI. 2026. *Introduction.* Retrieved 2026-09-24 from https://docs.typesafe.ai/introduction
[226] TypeSafe AI. 2026. *Jev 1.13 jaggedness.* Retrieved 2026-09-24 from https://docs.typesafe.ai/model-jaggedness/jev-1.13
[227] TypeSafe AI. 2026. *Models.* https://docs.typesafe.ai/models.md
[228] TypeSafe AI. 2026. *Patterns.* Retrieved 2026-09-24 from https://docs.typesafe.ai/patterns
[229] TypeSafe AI. 2026. *TypeSafe AI Home Page.* Retrieved 2026-09-24 from https://typesafe.ai/
[230] TypeSafe AI. 2026. *Workflow evals.* Retrieved 2026-09-24 from https://evals.typesafe.ai/
[231] Naman Vats. 2026. Good Judgment, Bad Advice: Decision-Model Controllers Raise Agent Cost Without Improving Success. arXiv:2610.24649 [cs.AI] https://arxiv.org/abs/2610.24649
[232] Wadhwa. 2026. *OpenDecision: An Open-Source Semantic Decision Engine.* Retrieved 2026-09-24 from https://github.com/deepanwadhwa/OpenDecision
[233] Naoki Wake and Justin Wagle. 2026. SharedKV-BT: Node-Local Typed Decisions for Behavior-Tree Agents. arXiv:2610.07327 [cs.RO] https://arxiv.org/abs/2610.07327
[234] Guangren Wang. 2026. Mnemon: Raw Records, Fast Judgments, Slow Thoughts. arXiv:2609.36059 [cs.CL] https://arxiv.org/abs/2609.36059
[235] Jianing Wang, Youwen Jian, Derong Wang, Jinyang Yu, Qingyi Si, Chuanyu Qin, Yanghua Xiao, and Hongcheng Guo. 2026. UniJev: Separating Accuracy from Calibration in Trimodal System One Decision Models. arXiv:2610.25784 [cs.CV] https://arxiv.org/abs/2610.25784
[236] Jianye Wang, Kexin Zhang, Jiayin Xiong, and Hao Xue. 2026. Benchmarking System One Models for Time-Series Decision Making: A Case Study of Jev. arXiv:2610.33410 [cs.AI] https://arxiv.org/abs/2610.33410
[237] Kaixiang Wang, Zhaijiacheng Zhou, Jiong Lou, and Jie Li. 2026. Jev-LiteMAS: Separating Decisions from Generation in Multi-Agent Workflows. arXiv:2610.21411 [cs.AI] https://arxiv.org/abs/2610.21411
[238] Shang Wang, Tianqing Zhu, Huajie Chen, Jiayang Li, Meng Yang, and Bo Liu. 2026. Hidden Risks of Jev: An Empirical Study of Security, Privacy, and Dual Use. arXiv:2610.04985 [cs.CR] https://arxiv.org/abs/2610.04985
[239] Xiuying Wang, Jiahua Cheng, Shuotian Li, Yufan Cheng, Bowen Deng, Zhexuan Bai, and Yichen Li. 2026. JevSoup: System-One Routing for Training-Free LoRA Composition. arXiv:2609.30922 [cs.AI] https://arxiv.org/abs/2609.30922
[240] Yuxiang Wang, Shengbo Cai, Zhixian Zhao, Yingda Shen, Yuancheng Wang, Xueyao Zhang, Liqiang Zhang, Teddy Sun, Steve Yves, and Zhizheng Wu. 2026. Jarvis: Coordination Deserves a Brain of Its Own in Agentic Interactive Models. arXiv:2610.23804 [cs.SD] https://arxiv.org/abs/2610.23804
[241] Yang Wang, Hanlin Xiao, Yanan Ma, Chenghao Xiao, and Chenghua Lin. 2026. Calibrated Typed Decisions for Entity, Relation, and Event Extraction. arXiv:2610.17113 [cs.CL] https://arxiv.org/abs/2610.17113
[242] Yuhao Wang and Zelin Zang. 2026. ABACUS: Program-Guided Multistep Reasoning over Frozen Judgment Models. arXiv:2610.25963 [cs.AI] https://arxiv.org/abs/2610.25963
[243] Zhiqiang Wang and Yichao Gao. 2026. JEV as a Judge for Agent Trace Security: An Empirical Comparison with Generative LLM Judges. arXiv:2609.34862 [cs.CR] https://arxiv.org/abs/2609.34862

[244] Zhongyin Wang, Zhaoqian Xue, Junting Mu, Haozheng Liu, Ye Yuan, Haolun Wu, and Lingyao Li. 2026. Truncated Reasoning Under a Budget: Evaluating System-1 Models as Decision-Makers. arXiv:2610.25768 [cs.CL] https://arxiv.org/abs/2610.25768

[245] Zhengle Wang, Hanxu Yan, Fuheng Zhao, and Chunwei Liu. 2026. Prune First, Decide Fast: Scalable Semantic Query Processing with JEVDB. arXiv:2610.02046 [cs.DB] https://arxiv.org/abs/2610.02046

[246] Zexiao Wang, Zihao Zhang, Xudong Wang, Pan Wang, Ziyi Ye, Haoyu Zhao, Zuxuan Wu, and Shuicheng Yan. 2026. Chinese-Jev: Bringing System One Model to Chinese-Language Tasks. arXiv:2609.36965 [cs.CL] https://arxiv.org/abs/2609.36965

[247] Yuchuan Weng, Chi Liu, Xin Chen, and Srinivasan Manoharan. 2026. Our Own Jev: Typed Decisions in One Forward Pass, from an In-House Judge to Open LoRA Models. arXiv:2610.29052 [cs.CL] https://arxiv.org/abs/2610.29052

[248] wfzyx. 2026. *wfzyx/von · Hugging Face*. Retrieved 2026-09-24 from https://huggingface.co/wfzyx/von

[249] Penny F. Whiting, Anne W.S. Rutjes, Marie E. Westwood, Susan Mallett, Jonathan J. Deeks, Johannes B. Reitsma, Mariska M.G. Leeflang, Jonathan A.C. Sterne, Patrick M.M. Bossuyt, and the QUADAS-2 Group*. 2011. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine* 155, 8 (2011), 529–536. doi:10.7326/0003-4819-155-8-201110180-00009

[250] Robert F. Wolff, Karel G.M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, Sue Mallett, and for the PROBAST Group†. 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine* 170, 1 (2019), 51–58. doi:10.7326/m18-1376

[251] Chenkai Wu, Xing Chang, and Jun Yan. 2026. Poisoned into Action: Context Attacks on Typed Decision Workflows. arXiv:2610.15537 [cs.CR] https://arxiv.org/abs/2610.15537

[252] Pengying Wu, Zhejia Wen, Kang Ding, Kangjie Zhou, Chunxuan Jiao, Wei Pan, Xiangli Nie, and Chang Liu. 2026. CovNav: Conformal Prediction-Guided Adaptive Inference for Object Goal Navigation. arXiv:2610.27677 [cs.RO] https://arxiv.org/abs/2610.27677

[253] Tiantong Wu and Wei Yang Bryan Lim. 2026. Decision Hijacking: Prompt Injection Attacks on Jev's Typed Probabilistic Decisions. arXiv:2609.28613 [cs.CR] https://arxiv.org/abs/2609.28613

[254] Tiantong Wu and Wei Yang Bryan Lim. 2026. REFLEX with Jev for Efficient Selective Control in LLM Agents. arXiv:2609.26532 [cs.AI] https://arxiv.org/abs/2609.26532

[255] Yuchao Wu. 2026. SAG2: Embedding-Free Multi-Hop Retrieval with BM25 and JEV. arXiv:2610.22588 [cs.CL] https://arxiv.org/abs/2610.22588

[256] Yuanbo Xie, Mingli Zhu, Tianyun Liu, Yingjie Zhang, Haixiao Xiong, and Tingwen Liu. 2026. Lost in the Fallback: How Reliably Do Typed Decision Models Use Fallback Options? arXiv:2610.19856 [cs.CR] https://arxiv.org/abs/2610.19856

[257] Chuhong Xu, Bo Su, Ziyao Chen, Ruiyang Xu, Shimeng Dai, and Xinyu Qiu. 2026. Same-Number Citation Swaps: Stress-Testing Jev as a Financial Evidence Judge. arXiv:2610.08675 [cs.CL] https://arxiv.org/abs/2610.08675

[258] Shenghe Xu and Lifan Mei. 2026. A First Glance at Jev for Network Traffic Classification: Accuracy, Processing Time, and Cost. arXiv:2610.00376 [cs.LG] https://arxiv.org/abs/2610.00376

[259] Zhiwei Xu, Xuyang Liu, Xiaolin Peng, Weixian Deng, Jiabao Gao, Zirui Li, Tian Qiu, Boya Zhou, Wei Jiao, Hai Wan, and Xibin Zhao. 2026. TrustJev: Separating Policy Authority from Untrusted Evidence in LLM-Based Decisions. arXiv:2610.34826 [cs.AI] https://arxiv.org/abs/2610.34826

[260] Zixiang Xu, Zirui Song, Chiyu Zhang, Xiuying Chen, Xi Liu, Xiyang Hu, and Yue Zhao. 2026. JevOut: Natural Context Can Flip Decision Models. arXiv:2609.30243 [cs.CL] https://arxiv.org/abs/2609.30243

[261] Karan Yadav. 2026. Two Uncertainties Are Not One: Disagreement versus Decision-Risk in Cheap Structured Content Moderation. arXiv:2610.33258 [cs.CL] https://arxiv.org/abs/2610.33258

[262] Kakei Yamamoto and So Kuroki. 2026. PRISM: Calibrating Black-Box Judges with Counterfactual Probes. arXiv:2610.18389 [cs.LG] https://arxiv.org/abs/2610.18389

[263] Qiang Yan. 2026. Chinese Decision: A Family of Language Models for Single-Pass Probabilistic Decision-Making. arXiv:2610.23170 [cs.CL] https://arxiv.org/abs/2610.23170

[264] Yu Yan. 2026. JevForest: Path Voting for Budgeted Feature Acquisition. arXiv:2610.10615 [stat.ML] https://arxiv.org/abs/2610.10615

[265] Wang Yang, Xinpeng Li, Yanyan Zhang, Chuang Ma, Shouren Wang, Jianzhe Wei, Liqin Ye, Xiaoran Cai, Mohsen Hariri, Vipin Chaudhary, and Xiaotian Han. 2026. What Can't Jev Decide? Probing the Limits of One-Pass Decision Models. arXiv:2610.24929 [cs.CL] https://arxiv.org/abs/2610.24929

[266] Xianliang Yang, Yapu Zhang, and Li Zhao. 2026. GraphDecisionBench: Benchmarking System One Models on Graph Tasks. arXiv:2610.06354 [cs.AI] https://arxiv.org/abs/2610.06354

[267] Yaodong Yang, Hongyao Tang, Yi Ma, Xingyu Fan, Weixun Wang, Jinpeng Li, and Tianpei Yang. 2026. Code Owns the Simulation, Jev Owns the Evaluation. arXiv:2610.01834 [cs.AI] https://arxiv.org/abs/2610.01834

[268] Yiqu Yang, Qiaoyuan Zheng, Yang Tian, Yu Fan, and Elliott Ash. 2026. TUPLE-Bench: Evaluating and Mitigating Privacy-Utility Trade-offs in Typed Decision Models. arXiv:2610.23668 [cs.CR] https://arxiv.org/abs/2610.23668

[269] Weiwei Ye, Hangchen Liu, and Renhe Jiang. 2026. NumericJev: Jev-like LLM Numerical Decoding with Multiway Decision Trees. arXiv:2609.28587 [stat.ML] https://arxiv.org/abs/2609.28587

[270] Furkan Yilmaz, Habibe Aleyna Tasdemir, and Muhammed Faruk Gozay. 2026. LAVOIR: Teaching a Single-Pass Decision Encoder When and What to Ask with Amortized Value of Information. arXiv:2609.30706 [cs.AI] https://arxiv.org/abs/2609.30706

[271] Liang You, Zhiyin Zhou, Dawei Tang, Chuchu Wu, and Jingzhuo Hu. 2026. Native Probability Outputs for Clinical Triage: An Evaluation of Calibration and Decision Accuracy. arXiv:2610.24692 [cs.LG] https://arxiv.org/abs/2610.24692

[272] Guanxu Yu and Yuhang Yao. 2026. Visual Jev: Accurate and Efficient Decisions from Shared Visual Context. arXiv:2609.25845 [cs.CV] https://arxiv.org/abs/2609.25845

[273] Yundaichuan Zhan, Weishi Wang, Wenbiao Liu, Daniel Dahlmeier, Chengwei Qin, Juncheng Li, Fredrik D. Johansson, and Zhongqi Yue. 2026. Dyad: Extending Large Language Models with Native Typed Decision-Making. arXiv:2609.36116 [cs.LG] https://arxiv.org/abs/2609.36116

[274] Andrew Zhang, Trinh T. L. Vuong, Sophia J. Wagner, Dongmin Bang, Tong Ding, Yuchang Su, Mary Cipperman, Christina S. Chen, Judith Lefkes, Ramin Kahidi, Ming Y. Lu, Caiwei Tian, Tianhao Luo, Jingnan Ma, Rowland Pettit, Long Phi Le, and Faisal Mahmood. 2026. A general-purpose System One decision model for clinical records. arXiv:2610.27099 [cs.AI] https://arxiv.org/abs/2610.27099

[275] Baowen Zhang, Hangting Ye, and Hongyuan Zha. 2026. Typed Decision Models on Tabular Data. arXiv:2610.19536 [cs.CL] https://arxiv.org/abs/2610.19536

[276] Fan Zhang, Yankai Chen, Zhuohan Xie, Yixi Zhou, Sijia Peng, Lei Fan, Xinhua Ji, Cunyuan Zheng, Huangyong Shan, Philip S. Yu, Xue Liu, Yu Chen, Preslav Nakov, and Songwei He. 2026. Same Scores, Different Decisions: Evaluating JEV and Language Models for Legal Document Understanding. arXiv:2609.27678 [cs.CL] https://arxiv.org/abs/2609.27678

[277] Jiamu Zhang, Tianze Yang, Yucheng Shi, Evan Chen, Zixiang Nie, Kelly Wan, Liangjie Hong, Ninghao Liu, and Liang Wu. 2026. AnyJev Technical Report. arXiv:2610.00831 [cs.LG] https://arxiv.org/abs/2610.00831

[278] Jiahao Zhang, Yifan Zhang, and Yu Huang. 2026. JevCoder: When Does Decision Delegation Pay Off in Code Repair? arXiv:2610.19596 [cs.SE] https://arxiv.org/abs/2610.19596

[279] Linghua Zhang. 2026. Jev-Mobile: Jev as an Executor for Mobile GUI Agents. arXiv:2609.30186 [cs.AI] https://arxiv.org/abs/2609.30186

[280] Yiqun Zhang, Peidong Wang, Zihan Wang, and Shi Feng. 2026. Decide, Don't Generate: Competitive Dimensional ABSA with Jev's Typed Decisions. arXiv:2609.35293 [cs.CL] https://arxiv.org/abs/2609.35293

[281] Yazhou Zhang and Junhao Yu. 2026. Emo-Jev: Probabilistic Reasoning for Emotion Classification with Jev. arXiv:2610.08829 [cs.CL] https://arxiv.org/abs/2610.08829

[282] Xinyu Zhao, Qingyun Sun, Ziwei Zhang, Tianchen Zhu, Daqing Li, and Jianxin Li. 2026. One Pretrained Decision Model Is Enough for Training-Free Generalist Graph Anomaly Detection. arXiv:2610.30777 [cs.LG] https://arxiv.org/abs/2610.30777

[283] Bin Zheng and Zheng Chen. 2026. Selective Reasoning for Typed Decision Models: When to Start, Read, and Stop. arXiv:2610.19644 [cs.CL] https://arxiv.org/abs/2610.19644

[284] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*. 46595–46623. arXiv:2306.05685 doi:10.52202/075280-2020

[285] Jike Zhong, Ming Li, and Yuxiang Lai. 2026. When the Right Answer Is Missing: An Arithmetic-Dependent Rejection Bottleneck in Jev. arXiv:2609.39496 [cs.LG] https://arxiv.org/abs/2609.39496

[286] Xunlan Zhou, Xianliang Yang, and Li Zhao. 2026. From Text Decisions to Pixels: An Study of Jev-Style Visual Choice Model. arXiv:2609.29283 [cs.AI] https://arxiv.org/abs/2609.29283

[287] Zihan Zhou, Xinzhe Hu, Hanxu Yang, Liangjian Wen, and Zhao Kang. 2026. Token-Efficient Multi-Agent Collaboration via System One-Guided Computational Division of Labor. arXiv:2610.08155 [cs.MA] https://arxiv.org/abs/2610.08155

[288] Kunlun Zhu, Tao Feng, Xu Han, Heng Wang, Ling Yang, Xiangru Tang, Jiaxuan You, and Heng Ji. 2026. JevAgentHarness: Benchmarking, Deploying and Training Decision Models That Share Control with LLMs. arXiv:2610.21412 [cs.AI] https://arxiv.org/abs/2610.21412

[289] Tong Zhu, Zhe Xu, Zefeng He, Yunkai Chai, Xuyang Hu, Dehong Zheng, Yun Luo, Yafu Li, Qi Zhang, and Yu Cheng. 2026. Intern-Decision: Efficient Multimodal Decision Making with Autoregressive Masked Language Modeling. arXiv:2610.22131 [cs.CL] https://arxiv.org/abs/2610.22131

[290] Yuhan Zhu, Xiangyu Zeng, Qingyi Si, Chenxu Yang, Yifei Li, Luozheng Qin, Junhao Zhou, Dingyu Yao, Jun Zhang, Chuanyu Qin, Zhiqiu Zhang, Yuandong Yang, Yuqi Xu, Tianyu Zhang, Hongcheng Guo, Jiaqi Wang, Nan Duan, and Limin Wang. 2026. VideoJev: System One Model for Video Understanding and Interaction. arXiv:2610.18615 [cs.CV] https://arxiv.org/abs/2610.18615

[291] Xialie Zhuang, Mingcong Lei, Zunhai Su, Ri Liu, Jianjin Li, Peixian Ma, Jia Pan, Ngai Wong, Guanhua Chen, and Chuan-Wei Kuo. 2026. DecisionPivot: A Diagnostic Benchmark for Decision Models in LLM Systems. arXiv:2610.28522 [cs.AI] https://arxiv.org/abs/2610.28522

[292] Xialie Zhuang, Zunhai Su, Jia Pan, Ngai Wong, Guanhua Chen, and Chuan-Wei Kuo. 2026. StartLux-Decision: Auto-Think Decision Models with Reinforcement Learning for Calibrated Deliberative Decisions. arXiv:2610.17309 [cs.CL] https://arxiv.org/abs/2610.17309