# 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

Typed probabilistic decision models, which their first vendor calls System One models, return a probability distribution over a caller-declared answer space and promise that it is calibrated. They form a contract assembled from five research lines. The survey provides a genealogy, taxonomy, formal core, ecosystem census, graded evidence ledger, benchmark summary, pattern catalog with failure atlas, and reporting standard. The evidence supports three regularities. The contract travels, while robustness to option names belongs to each checkpoint. Calibration belongs to the model and the task together. Escalation earns the saving when confidence ranks errors and the price gap is wide.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Natural language processing**; Learning paradigms; • **General and reference** → *Surveys and overviews.*

Additional Key Words and Phrases: System One models, decision models, typed probabilistic decision models, Jev, confidence calibration, selective prediction, learning to defer, LLM cascades, LLM routing, zero-shot classification, dual-process AI, reporting standards

## 1 Introduction

A program that asks a model for a decision needs one of the options it declared and a number that it can compare with a threshold. A generative model returns text instead, so the program must parse the answer, validate it, and elicit a confidence that may not track correctness [140, 163]. A typed probabilistic decision model returns a probability distribution over a caller-declared finite answer space in one forward pass, and it promises that the distribution is calibrated (Definition 2.1). The contract became a product on 15 September 2026, when TypeSafe released the hosted model Jev under the pinned version jev-1.13.0 and called it a System One model [147, 153]. Open implementations on public backbones followed within days [8, 25, 116], so a program could obtain the same contract from several models at once.

### 1.1 Motivation for a survey of the contract rather than the product

The object of this survey is the contract rather than any single product, and the dispute over the class name in the week of the launch shows why. The vendor wrote System One model, commentators and evaluators wrote decision model [58, 160], and authors probing one interface behavior wrote typed decision model [16, 139]. This survey defines the class as a typed probabilistic decision model, uses decision model as its working term, and keeps System One model for the vendor's category and the dual-process lineage. Section 2 gives the reasons for that choice, together with the scope, search, evidence roles and boundaries of the review.

The thesis of this survey is that decision models are a new contract rather than a new kind of intelligence. Five research streams converge on that contract, as Figure 1 shows, with milestones dated from the works of Section 3, Sentence-BERT [125] and language-model cascades [32]. Label-conditioned classification supplies the scoring head, reward models and judges the verdict role, proper scoring rules the training target, the reject option the policy that acts on a probability, and dual-process artificial intelligence the fast-solver role and the name. September 2026 added a stable interface of three primitives, calibration stated as a product promise, a price billed per input token,

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.

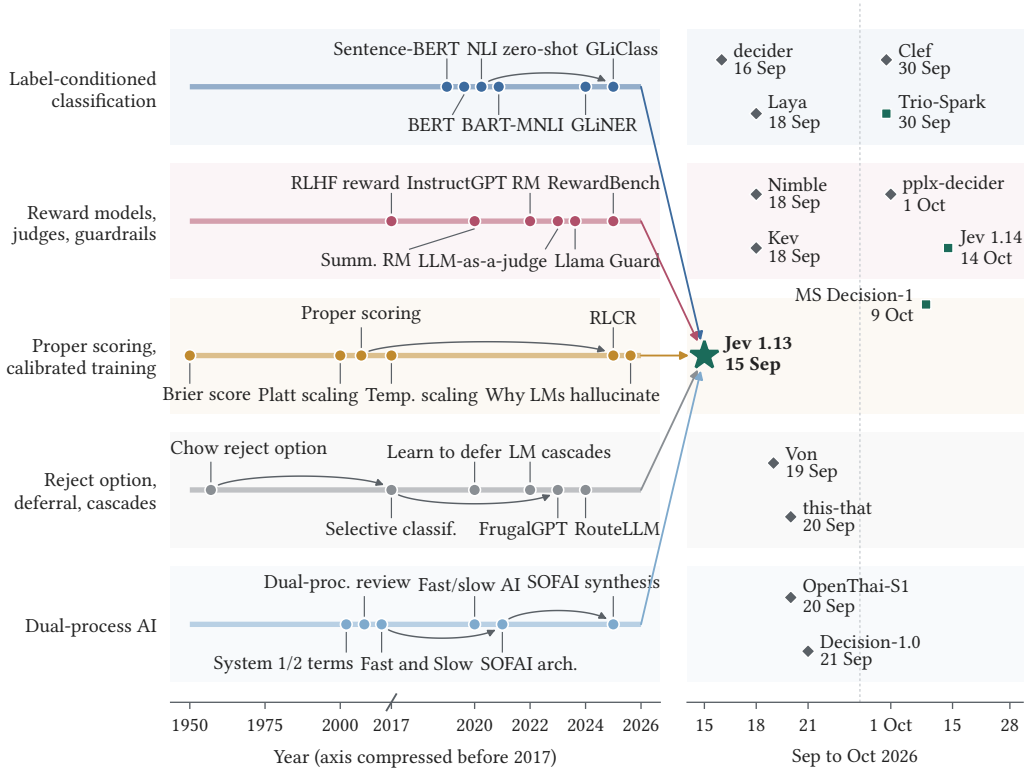



**Fig. 1.** Five research lines on a compressed dated axis converge on the launch of jev-1.13.0 on 15 September 2026. The right strip shows the decision-model releases to 28 October 2026, open ones as diamonds dated from their census cards and hosted ones as squares dated from their vendor pages, with the first cutoff dotted.

and an open ecosystem of replicas. The unit of evaluation is therefore the contract together with the policy that consumes it.

The evidence gathered to the cutoff supports three regularities, stated in the form that the synthesis of Section 7 can defend. The contract travels, and robustness to option names belongs to each checkpoint, so it must be tested per model. Open heads reproduced the interface within days. For open models used zero-shot, the median accuracy gap to the hosted model, open minus hosted, was −0.7 points across 35 studies, while models trained on the target task led it. Under one harness, an option-name swap changed most answers of the hosted model and Laya and few of the other open families (Section 8). Calibration belongs to the model and the task together, so it is audited per schema. The hosted model's raw calibration error varied by a factor of 77 across studies under different estimators, and from 0.063 to 0.538 across tasks within one study [47, 58, 122]. Escalation earns the saving when confidence ranks errors and the price gap is wide. Across the 23 cascade studies of Section 7.3, confidence-gated escalation kept a median 0.99 of a stronger model's quality at a median 0.56 of its fee, under cost definitions that differ by study [for example 58, 81, 162]. The saving shrank as the price gap narrowed, and a cheap generative cascade matched it where routing was already accurate [58, 162].

### 1.2 Contributions and organization

The contributions of this survey follow the order of its sections, and each serves either a program that consumes the contract or a study that evaluates it. Section 3 gives the dated genealogy that Figure 1 summarizes, and it shows what each of the five streams supplied to the contract. A taxonomy of decision heads, training objectives and interface contracts follows in Section 4, with the head families in Figure 3. Section 5 states six propositions on isolation, resolution, confidence and cascades that hold for any model honoring the contract, under the premises each one states. A claim about one model can then be checked against what the contract alone implies. The census of models, datasets and harnesses in Section 6 counts 101 original implementations in 67 recipe families at the cutoff of 24 September 2026. Section 7 grades each of the 248 studies in an evidence ledger for risk of bias, records its evidence role, and synthesizes the ledger into the three regularities stated above.

The remaining contributions carry the synthesis into measurement and practice. Section 8 summarizes a same-harness measurement of the hosted model and several open decision models, reported in the companion paper [121], with the same semantic request for every model and conventional baselines beside them. Section 9 fuses a catalog of design patterns with an atlas of failure modes, and Section 10 reviews applications by evidence grade. Section 11 proposes a reporting standard, and Section 12 sets out an agenda of open problems with criteria that a future study can pass or fail.

Readers can follow one of two paths through these sections, according to whether they build with decision models or study them. Practitioners who build software on decision models can read Section 4 for the interface, Section 9 for the patterns and failure modes, and Section 11 for the checks that a deployment reports. Researchers in calibration and selective prediction can read Section 5, Section 7 and Section 12, which carry the propositions, the graded evidence and the open problems. Readers new to the area can begin with Section 3, which places each stream before the product, and Section 13 draws the paths together. The electronic supplement holds the proofs, search log, risk-of-bias ratings, full checklist, census, pattern entries and ledger by study, and the main text points to each part where it is used.

## 2 Scope, method and terminology

This section fixes the object of the survey, how its evidence was gathered and graded, and the terms used throughout. The class is stated formally in Definition 2.1, and the remaining paragraphs build on it.

**Definition 2.1 (Typed probabilistic decision model)**

Given a state $x$ collecting the evidence available to a request, and a question $q$ declaring a finite answer space $\mathcal{Y}_q$ of size $K$ with the criteria fixing each answer's meaning, a typed probabilistic decision model is a mapping

$$(x, q) \longmapsto \big(\hat{p}(\cdot \mid x, q),\ c(\hat{p})\big), \qquad \hat{p}(\cdot \mid x, q) \in \Delta(\mathcal{Y}_q), \tag{1}$$

returned in one forward pass without generating unconstrained text, where $\Delta(\mathcal{Y}_q)$ is the simplex on $\mathcal{Y}_q$ and $c(\hat{p})$ is an optional confidence statistic from $\hat{p}$, which Noul questions omit. The contract promises that $\hat{p}$ is calibrated in a sense that Section 5 makes precise, and the model honors question isolation when $\hat{p}(\cdot \mid x, q)$ depends on $x$ and $q$ alone, so several questions in one request return separate marginals rather than a joint distribution.

The three documented question primitives, Choice, Score and Noul, instantiate this template, and Section 4 describes each [144, 148, 152]. The confidence statistic on Choice and Score answers is a separate returned quantity that we audit rather than assume meaningful [146]. Section 5 builds this notation into propositions on isolation, resolution, confidence and cascades.

The survey includes a source that reports a measured quantity, a documented interface, or a governing method for a model meeting Definition 2.1. It excludes tutorials with no measurement and search-oriented directories, which the census lists without treating them as evidence. The ledger admits evaluations of models released under the contract from the launch on, while older members of the class, such as reward models and zero-shot classifiers, enter through the genealogy of Section 3. Sources were the arXiv application programming interface, citation chasing from the launch post and two early evaluations through OpenAlex and Semantic Scholar, Hugging Face and GitHub search, and a curated list of web pages. A first sweep ran 23 arXiv queries to a cutoff of 24 September 2026, and a second ran 24 queries, adding the names of models released later, over submissions to 28 October 2026. Together they returned 271 candidates, included 243 against Definition 2.1 and excluded 28. A further 3 included papers reported no measurement at extraction, and with 8 grey-literature studies the ledger of Section 7 holds 248 studies. Section S2 lists every query and outcome, including the first sweep's 513 Hugging Face hits and 226 GitHub repositories.

This survey organizes every extracted claim by four evidence roles. Methodological research supports general claims about learning, calibration, inference or evaluation. Product documentation states the advertised interface and its disclosed constraints. Vendor evaluation reports a measurement under a provider's chosen tasks and comparisons. Application research evaluates a model inside a pipeline and inherits its choices. These roles are not a trustworthiness ranking, since a documented interface can be strong evidence only for itself. The ledger of Section 7 labels each row with one of three estimands, namely agreement with another model, agreement with an administrative field, or correctness against an adjudicated label. It also states the estimands behind every summary that pools them.

A public companion repository at https://github.com/pozapas/system-one-models-survey holds the census, the evidence ledger and the harness behind the measurements of Section 8, each row carrying a source URL and an access date. Its census and ledger are versioned rather than edited in place, so a correction after the cutoff becomes a new row that supersedes the old one. This survey describes version v1.2.3, which adds the second sweep, its extraction, its grading and its adjudicated second rating to version v1.1, the release of the first cutoff. Any later sweep is released as a new minor version with its own recorded date. A public curated index, maintained alongside this survey at https://github.com/pozapas/awesome-system-one-models, lists the papers, models, datasets, harnesses and graded studies covered here and is updated after the cutoff.

Three names compete for the same object, System One model, decision model and typed decision model, and Table 1 states how this survey uses each. We define the class as a typed probabilistic decision model in Definition 2.1 and adopt decision model as the working term. System One model is reserved for the vendor's category and the dual-process lineage, whose history Section 3.5 traces.

Table 2 places this survey against the closest reviews. Shorinwa et al., Liu et al. and Geng et al. survey uncertainty quantification, calibration and confidence elicitation for models that generate text [44, 87, 133]. Strong et al. survey learning to defer, Varangot-Reille et al. with Moslem and Kelleher survey routing and cascading, and Hendrickx et al. survey the reject option generally [52, 104, 138, 155]. Each covers one genealogy stream of Section 3 rather than the interface, the ecosystem or the evidence assembled here. Three reviews of the class itself appeared after the launch. Tang and Zheng audit 28 papers posted in the first nine days and derive an evaluation checklist [141]. Their reading of the readout, cost and escalation evidence agrees with ours, but 27 of those papers are also in our ledger, so the agreement is a shared reading of shared evidence rather

**Table 1.** Terminology used in this survey. Bold marks the working term.

| Term | Role in this survey | Usage elsewhere |
|---|---|---|
| Typed probabilistic decision model | Formal name of the class, given in Definition 2.1 | Introduced here |
| **Decision model** | Working term used through the remaining sections | Ibrahim and Zaki [58], Willison, Simon [160] |
| System One model | Reserved for the vendor's product category and the dual-process lineage | Booch et al. [10], TypeSafe AI [153] |
| Typed decision model | Synonym in independent evaluations, not adopted here | Cheng et al. [16], Sun et al. [139] |

**Table 2.** Relation to prior surveys. A filled dot marks each genealogy stream of Section 3 or survey layer that a review treats as its object, and a half dot one it treats in part.

| Survey and scope | Genealogy stream | | | | | Survey layer | | |
|---|---|---|---|---|---|---|---|---|
| | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 | Contract | Census | Ledger |
| Shorinwa et al. [133]; Liu et al. [87] Uncertainty and calibration of generative language models | ○ | ○ | ● | ○ | ○ | ○ | ○ | ○ |
| Geng et al. [44] Confidence elicitation in language models | ○ | ○ | ● | ○ | ○ | ○ | ○ | ○ |
| Strong et al. [138] Learning to defer to an expert | ○ | ○ | ○ | ● | ○ | ○ | ○ | ○ |
| Varangot-Reille et al. [155]; Moslem and Kelleher [104] Routing and cascading among language models | ○ | ○ | ○ | ● | ○ | ○ | ○ | ○ |
| Hendrickx et al. [52] The reject option in machine learning | ○ | ○ | ○ | ● | ○ | ○ | ○ | ○ |
| CACM and npj AI syntheses [7, 40] Dual-process, fast-and-slow architectures | ○ | ○ | ○ | ○ | ● | ○ | ○ | ○ |
| Tang and Zheng [141] Early evidence audit of 28 decision-model papers with an evaluation checklist | ◑ | ◑ | ◑ | ◑ | ○ | ◑ | ◑ | ◑ |
| Li et al. [77] Semantic decision engines in network control loops | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◑ |
| Roitman [130] Practitioner guide to fast decision models in five families | ◑ | ● | ● | ◑ | ● | ◑ | ◑ | ○ |
| This survey The typed, calibrated, non-generative decision contract | ● | ● | ● | ● | ● | ● | ● | ● |

than an independent replication. Li et al. systematize semantic decision engines in network control loops and show that a decision meeting its time budget in isolation can miss it once requests queue [77]. That result qualifies latency claims measured one request at a time. Roitman's practitioner guide sorts fast decision models into five families and treats calibration and cascades at length, without a stated search or a graded ledger [130]. None of the three states the contract formally with propositions, grades the risk of bias of its studies, or ties each number to a verbatim source, and these are the layers this survey adds. A concurrent study that appeared after the sweep analyzes public GitHub projects that call the hosted model and how they combine its primitives [85]. It complements the census of Section 6 with a view of use rather than of implementations. The 2025 Communications of the ACM and npj Artificial Intelligence syntheses treat dual-process, fast-and-slow architectures as a cognitive framing, and we draw on their history in Section 3.5 while keeping its object narrower [7, 40].

We set the inclusion rules, the risk-of-bias domains and the synthesis protocol before extraction. We then ran the arXiv, Hugging Face and GitHub sweeps, extracted the census cards and ledger rows with a verbatim excerpt attached to every extracted number, and checked each citation against its registry record. For the first sweep, a second rater outside the author team independently repeated

the screening and a random sample of the extraction and of the grading. A further 5 raters graded the remaining studies and read one value from each. For the second sweep, 10 second raters who never saw the first rating repeated the screening of every candidate, the grading of every study, the extraction of at least one value per study and a sample of the head codes. The studies added at adjudication were rated by 3 more, so every graded study has an independent second rating. We adjudicated every disagreement against the full text, as Section S3 reports. We graded every study of the second sweep against its full text rather than from the extraction note alone. Every extracted excerpt was also checked programmatically against the saved full text and against its extracted value.

## 3 Genealogy of the decision contract

The decision contract of Definition 2.1 is best read as a convergence of five older research streams, not as an invention of September 2026. Label-conditioned classification supplies the scoring head, and reward models and judges supply the verdict role. Proper scoring rules give the training target, the reject option the policy that acts on the probability, and dual-process architectures the fast-solver role and the name. Figure 1 places the five streams on a dated timeline with one lane per stream.

### 3.1 Label-conditioned scoring heads

Classification over a fixed label set is the oldest ancestor of the contract. The contrast between generative and discriminative classifiers predates neural encoders [107], and pretrained bidirectional encoders made one forward pass the standard route to a class distribution [31]. A modern encoder of this family and its multilingual counterpart are the backbones of the two Laya checkpoints [24, 25, 157].

The decisive step moved the label from the output layer into the input. Entailment-based zero-shot classification recast each candidate label as a hypothesis scored against the text [76, 165], and generative models were prompted to score label descriptions [12]. Pattern-exploiting training phrased classification as cloze completion over verbalized labels [132], and GLiNER and GLiClass later encoded a text together with its label set [136, 168]. Once labels are read as text, their wording becomes part of the input. Predictions shift with label names and prompt format [171], and multiple-choice selection depends on option order and option symbols [128, 172]. Studies of language models as classifiers weigh these effects against fine-tuned encoders [174], and one evaluation reports that exchanging polar option names changed many more decisions than exchanging neutral names [139].

Every open decision model in the census except the fixed-slot head of Section 4 reads its options as text inside the context and scores them there. Laya scores each option at its own marker token [25], Kev reads option logits through a pointer head on a LoRA-adapted Qwen backbone [116], and Nimble and the decider family read answer-token logits [8, 97]. One open model describes its backbone as ModernBERT combined with GLiClass [51]. Because the hosted model's head is undisclosed, the label-in-context mechanism is documented for the open models alone. A black-box probing essay infers shared state encoding with isolated question branches and reports option-order sensitivity [57]. The chief executive endorsed a description of Jev as a zero-shot classifier [48], yet neither source identifies a head family.

Constrained decoding, structured output and function calling sit beside this stream rather than upstream of it. Grammar and schema constraints make a generator emit a valid type [33, 159], and tool-use training teaches it to emit calls in a declared format [131]. These methods get type validity from a sampled sequence, and format restrictions can change accuracy [140], so they are the natural comparator rather than a stream.

### 3.2 Reward models, verifiers, judges and guardrail classifiers

Reward models are decision models that were built and deployed before the name existed. Learning from human preferences trained a network to score one output against another [19, 137], and instruction-following pipelines made such a scorer a standard component of language-model training [110]. The scorer maps a state, here a prompt with a pair of candidate responses, to the probability that one is preferred, which is the contract of a binary decision model. Benchmarks now evaluate reward models as classifiers that separate preferred from rejected responses [73].

Verifiers carried the same idea from preference to correctness. A verifier trained to judge candidate solutions to math word problems reranked samples drawn from a generator [21], and process supervision scored each reasoning step rather than only the final answer [83]. Each step verdict is a binary decision with a probability attached. Pairing a producer with a separate scorer is an early form of the division of labor that decision-model cascades later adopt.

The LLM-as-a-judge literature made the verdict role explicit and exposed its biases. Strong language models were studied as substitutes for human evaluators on open-ended questions [173], and rubric-based scoring was formalized [89]. This line also documents position bias, self-preference, length bias and a wider taxonomy of judging errors [35, 117, 164, 173]. These biases are the judging counterparts of the option-order and label-name effects of Section 3.1, and they recur whenever a model reads the candidates it must choose among.

Guardrail and moderation classifiers form the production branch of the stream. Moderation classifiers screen undesired content at platform scale [98]. Llama Guard and ShieldGemma recast safety screening as instruction-conditioned classification over a declared taxonomy of harms [59, 169], which is a typed question in all but name. Recent evaluations of Jev as a judge connect the stream to decision models directly. One uses Jev as a judge that accepts when confident and escalates when unsure [81], and another evaluates Jev as a judge of clinical factuality in radiology reports [56]. This stream gives decision models a deployment role and a catalog of the ways in which presentation rather than content can sway a verdict model.

### 3.3 Proper scoring rules and calibrated training

Proper scoring rules supply both the training target of a decision model and the meaning of its promise. Brier introduced the quadratic score [11], Epstein extended scoring to ranked categories [38], Murphy partitioned the score into reliability, resolution and uncertainty [106], and the general theory of strictly proper rules followed [45]. Only the true conditional distribution optimizes such a rule in expectation, so training against one targets calibration in expectation, although finite data and a serving temperature can still leave a trained model miscalibrated.

Machine learning adopted the target through post hoc calibration and its measurement. Platt scaling, isotonic regression and Dirichlet maps turned classifier scores into probabilities [70, 120, 167]. Modern networks were found to be overconfident, and a single temperature became the default repair [46], a finding later revisited across architectures [101] and for pretrained transformers [30]. Calibration was also shown to degrade under dataset shift [111], and the binned expected calibration error [113] became the common summary, and later work examined how to estimate it [71].

Large language models brought confidence back as an explicit output. Models were shown to report elicited confidence of varying quality [63, 163], and linguistic calibration trained stated confidence against outcomes [5]. Calibration then entered the reward itself. RLCR augments a binary correctness reward with a Brier score on the stated confidence [27]. The hallucination argument makes a finite answer space attractive, since calibrated generators must hallucinate on rare facts and grading that rewards guessing sustains the behavior [65, 66].

Open decision models make the convergence concrete, since cross-entropy is the log score. Kev and Nimble train with cross-entropy over the declared options [8, 116], and the decider card adds policy-gradient training with a proper log score [97]. Laya names its recipe RLCD and rewards the log and spherical scores plus the ranked probability score for ordinal questions, optimized by REINFORCE with a group-mean baseline [25]. Its author presents a March 2025 arXiv paper as the start of this work [93] and also wrote a paper of 23 September 2025 on confidence-aware routing [92].

All that is known about how Jev was trained comes from the vendor's own statements. The launch post names Reinforcement Learning for Calibrated Decisions alongside a new architecture and a parallel sampler [147], and the primer presents RLCD as post-training toward probabilities that track correctness [143]. The chief executive described it as unpublished [75], so any reading of Jev's loss, reward, data or architecture, including the proper-score reading of an independent essay, is inference [57].

### 3.4 The reject option and its descendants

The reject option supplies the policy that turns a returned probability into an action. Chow described a recognizer that withholds a decision when its evidence is weak [18] and later derived the optimal tradeoff between error and rejection [17]. A statistical theory of learning with rejection followed [26], and selective classification turned the tradeoff into a risk-coverage curve with guarantees for deep networks [36, 43]. The top softmax probability became the baseline detector of misclassified and out-of-distribution inputs [53]. Cost-sensitive learning showed that the optimal threshold follows from the misclassification costs rather than from a fixed confidence level [37], and a caller applies the same rule to a decision model's probability.

Learning to defer replaced a fixed rejection with a learned choice between the model and a human expert. The model is trained jointly with a deferral rule that accounts for the expert's error [96], consistent surrogate losses give the rule statistical footing [105], and a recent survey organizes the field [138]. Conformal prediction adds distribution-free guarantees on top of any score. Prediction sets cover the true label at a chosen rate [2], risk-controlling sets and conformal risk control bound other losses [3, 6], and prediction sets have been applied to multiple-choice answers of language models [72].

Evaluating such a policy needs labels mainly where the policy is uncertain, which links deferral to label-efficient testing. Active testing estimates the risk of a fixed model from a small, deliberately chosen sample [69], and inverse-probability weighting keeps estimates unbiased when records are sampled with unequal probabilities [54]. Together these tools make a human review budget a quantity that can be estimated rather than guessed.

Cascades generalize deferral to a chain of models ordered by cost. The boosted detector cascade rejected easy negatives with cheap early stages [156]. Among language models, cascades escalate a query from a cheap model to a stronger one when a confidence signal falls below a threshold [14, 62], and routers learn the assignment from preference data [108]. Surveys of routing and cascading now organize this body of work [104, 155]. Decision models take from this stream a policy that consumes the probability. The stream also raises the question that Section 7 examines, which is when an escalation policy recovers a stronger model's quality at a fraction of its cost.

### 3.5 Dual-process architectures and the System One name

The name System One comes from dual-process accounts of human reasoning and judgment. Psychologists distinguished a fast, associative mode from a slow, rule-based one [39], and the labels System 1 and System 2 entered the reasoning literature through work on individual differences

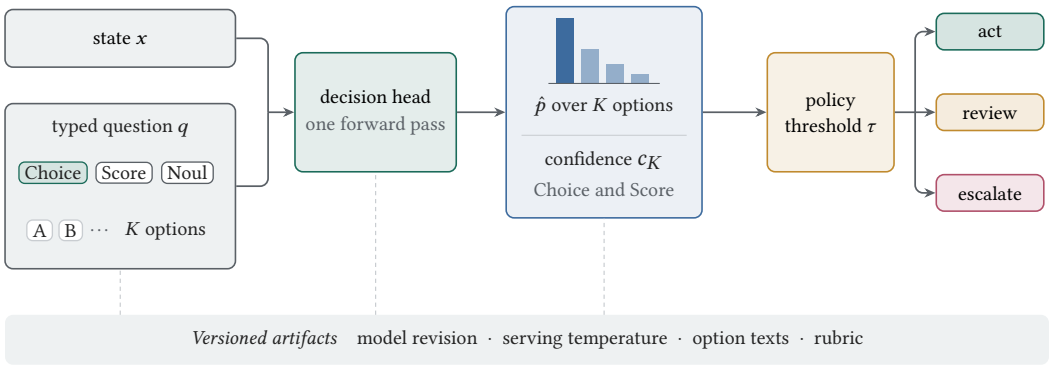



**Fig. 2.** The decision contract and the policy that consumes it. A decision head maps a state and a typed question to a distribution over the declared options, on which a policy acts, requests review or escalates.

[135]. Kahneman's popular synthesis later carried the terms from the reasoning literature to a wide audience [64].

Artificial intelligence research took up the framing as an architectural program well before September 2026. Booch and colleagues proposed thinking fast and slow as a research agenda for the field [10]. The SOFAI architecture built on it with fast solvers that answer from experience, slow solvers that deliberate, and a metacognitive component that governs which solver acts [41]. Syntheses in Communications of the ACM and in npj Artificial Intelligence later reviewed fast, slow and metacognitive thinking in machines [7, 40]. SwiftSage paired a fast small model with a slower planner for interactive tasks [84], and psychological studies examined intuitive behavior in large language models [49].

Before September 2026 the name therefore denoted a role inside a hybrid architecture rather than a class of models. In these designs the fast solver answers cheaply, reports how far its answer can be trusted, and hands off to a slower solver or a person under metacognitive control. The vendor's concept page takes the name from Kahneman's book and adds that confidence lets software decide when to act and when to escalate to a person or a reasoning model [153]. Laya's act and escalate head builds the same handoff into the model [25], and the cascades of Section 3.4 make the role operational.

The older name imports a cognitive claim that the models do not support, since nothing in the contract requires intuition, associative memory or a dual-process mind. Prior usage in artificial intelligence also named one component of a hybrid system rather than a product [10, 41], so we keep System One model for the vendor's category and this lineage. Commentators adopted the term decision model shortly after the launch [160]. Our formal class, the typed probabilistic decision model of Definition 2.1, makes no claim about how the model reasons.

## 4 Anatomy of a decision model

A decision model earns its calibration promise only together with the artifacts that define the question it was asked. Following the pipeline of Figure 2, we move from the versioned contract through the output head and the training objective to the latency and cost profile. Section 5 then states the properties that this anatomy makes possible.

### 4.1 The contract and its primitives

The effective predictor a caller receives is a bundle of versioned artifacts, not the network weights alone. Changing any one of them changes what is being evaluated, even when the billed model name stays the same. The hosted contract names a pinned model version rather than a moving alias. An alias that resolves to the newest release today can resolve to a different one tomorrow, while jev-1.13.0 keeps returning the behavior a threshold was tuned against [148]. This happened on 14 October 2026, when jev-latest moved from jev-1.13.0 to jev-1.14.0 and every unpinned caller changed model without changing a line of code [150]. The contract also fixes the schema of declared questions and answer spaces, the option texts or rubric levels naming each admissible answer, and the serving-time temperature applied before probabilities reach the caller. For calibration, two deployments that share a model version but differ in any of these artifacts are different predictors, because the mapping from a state and a question to a distribution depends on all of them jointly.

Three primitives carry these artifacts into a request. Choice supplies option names mapped to rubric descriptions, up to 255 options, and returns a probability per option plus the confidence statistic discussed below [144, 148]. Score supplies ordered level descriptions and returns a probability per level along with a single score, the probability-weighted mean of the level numbers [152]. Proposition 5.6 shows that this score cannot stand in for the cumulative probabilities that a threshold needs. Noul asks only for the probability that a yes-or-no statement holds, with no confidence value, since the vendor defines confidence only for the other two primitives [146].

None of the three primitives distinguishes an out-of-scope input from a plainly negative finding or from a request that was never properly evaluated. The vendor advises adding an option such as none of the above when a Choice list might not cover every case, which makes scope completeness a schema-design choice rather than a guarantee [148]. A contract reporting one probability per option also leaves the caller to decide whether its workflow needs top-label or classwise calibration, the two calibration targets that Section 5 defines formally. Top-label calibration can hold for the returned answer while classwise calibration fails on a rare class [46, 70].

### 4.2 Heads and objectives

Open implementations in the census realize the same contract with several recognizable families of output head. A fixed-slot head projects the state to a predetermined number of output positions and reads the answer from those logits, as OpenThai-SystemOne does with one slot reserved for abstaining. For a comparable head, accuracy falls sharply on more options than it saw in training [166]. In Kev-9B and Laya, a pointer or option-marker head reads the answer from a hidden state tied to each option’s own marker token. A single-token label-logit head reuses rows of the language-model head as option scores, softmaxed over the valid tokens only, as in decider-2b and Bespoke-Nimble-9B. The pngwn System One scorer uses a pair-scoring head, which scores each state-question-option triple as an inference model scores a premise against a hypothesis. A fifth family constrains the decoding of a generative backbone onto grammar-valid output, and no census model populates it. Label logits and option markers are the most common heads, in 171 and 135 of the 379 originals, while the heads of Jev and the other hosted models are undisclosed. The vendor notes only that high-cardinality Choice runs a two-stage process, scoring options independently before an explicit choice blamed for occasional slowdown [147]. A probing essay’s reading of shared-state encoding is an inference from latency, not an identified head [57].

The training objective shapes calibration once a head ships. Cross-entropy on fixed labels drives decider-2b’s supervised stages and open-jev-deberta-v3-large’s combined cross-entropy and Brier loss, which target label agreement rather than calibration and hence need a post hoc temperature fit. A proper-scoring-rule reward instead optimizes the reported distribution itself, as in Laya’s RLCD

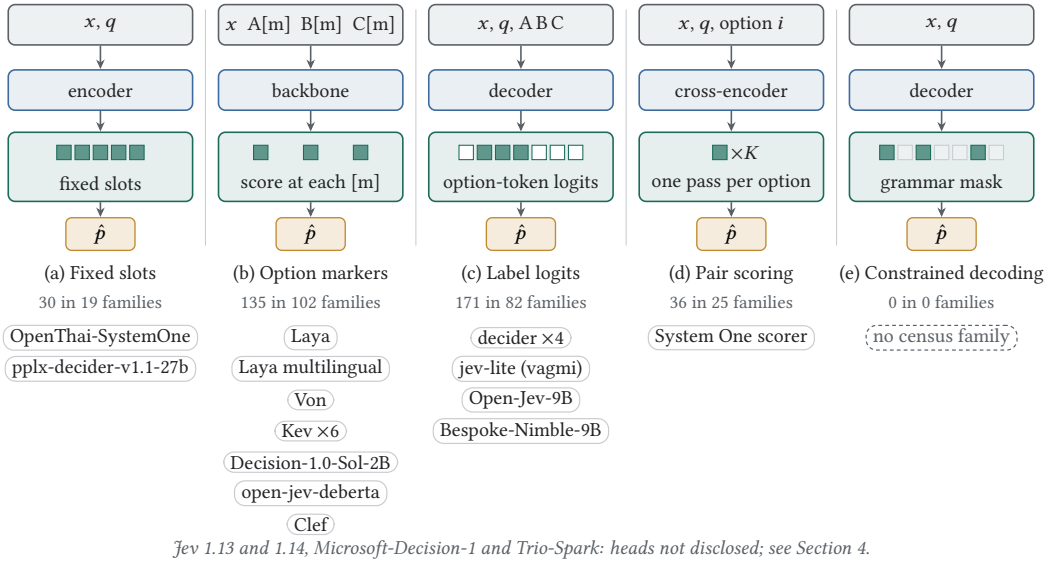



**Fig. 3.** Decision-head types of the census of Section 6. Under each panel, a count gives the originals and families of that type among all 379 originals, each coded from its own card or code. Chips name the families with an independent card, and undisclosed hosted heads are assigned to no panel.

stage and in the decider family's later stage, which adds a consistency term without further gold labels. Distillation transfers a teacher's soft label, and the jev-lite card trains against option-letter probabilities from a larger teacher while reversing option order to cancel position bias. A fourth path trains on contrastive pairs of synthetic decisions that change one fact so the correct answer flips, as in Bespoke-Nimble-9B's held-out pairs [9]. None ships a calibrated model by construction, and the fitted temperature differs by family. Kev and the decider family ship a temperature above their unscaled default, while Bespoke-Nimble-9B's newest checkpoint serves at that default and warns against reusing the older checkpoint's fitted value.

### 4.3 Confidence, isolation and cost

Every question is evaluated against the same state independently of the others, a design the vendor states directly [148]. Question isolation makes parallel evaluation practical, and Proposition 5.1 states what it withholds from a program that combines several answers. The contract is least precise about how it computes the confidence statistic attached to Choice and Score answers. The vendor's confidence page declines to publish a production formula and defers it to an unreleased cookbook. Only the page's code gives the normalized-maximum expression that reproduces its demonstration, and the code calls it an approximation for the three-option case [146]. A worked example on the jaggedness page does not reproduce that expression exactly [149], which is why Proposition 5.3 is stated for a class of statistics. Decider-2b states the same normalized-maximum expression as its own definition and ships it as production confidence for Choice. Laya's card adds an act-or-escalate output trained alongside its option-marker scorer, and the card finds that its probability carries almost no separating signal.

The state is encoded once, and every question is scored against that shared encoding rather than a separate copy. This is why the vendor describes adding questions as barely changing response time, and why independent probing found server time rising much more slowly than the number

**Table 3.** Notation of the formal core, with the result that defines each symbol.

| Symbol | Meaning | Defined in |
|---|---|---|
| $x$, $q$, $\mathcal{Y}_q$, $K$ | State, question, declared answer space and its size | Def. 2.1 |
| $\hat{p}$, $\hat{p}_{\max}$, $\hat{F}_k$ | Returned distribution, top probability, and cumulative probability up to level $k$ | Def. 2.1, Prop. 5.6 |
| $\delta$, $\rho$, $w_0$ | Smallest positive reportable probability, base rate, and share of zero reports | Prop. 5.2 |
| $c_K$, $\tau$ | Confidence statistic for $K$ answers and a threshold on it | Prop. 5.3 |
| $c_1$, $c_2$, $L_{\text{res}}$ | Mean costs of the two stages and residual loss of automated decisions | Prop. 5.4 |
| $c_{\text{FP}}$, $c_{\text{FN}}$, $c_{\text{R}}$ | Costs of a false positive, a false negative and a human review | Prop. 5.4 |
| $\alpha$, $\beta$, $\gamma$ | Rubric term, name term and interaction term of an option logit | Prop. 5.5 |

of questions asked [57, 148]. An open reimplementation sharing a visual backbone measured the mechanism directly. Reusing a cached prefix with batching lowered its per-question latency by 8.9 times over serial execution without reuse and reached 176 questions per second [166]. Jev is billed per input token, with output tokens unmetered because the contract returns no generated text, so one more question costs a few input tokens rather than another full request [148]. Total deployment cost over $N$ cases extends the per-item cost of Proposition 5.4. It adds a schema's setup cost $C_{\text{setup}}$, the inference cost $c_{\text{infer}}$ per case, and the escalation and review costs of the $N_{\text{esc}}$ escalated and $N_{\text{R}}$ deferred cases. It also adds the residual loss of automated decisions and the cost of monitoring and revising the schema,

$$C_{\text{total}} = C_{\text{setup}} + N c_{\text{infer}} + N_{\text{esc}} c_2 + N_{\text{R}} c_{\text{R}} + N\mathbb{E}[L_{\text{res}}] + C_{\text{monitor}} + C_{\text{revision}}. \tag{2}$$

A low price per input token bounds only the second term of this expression, so a deployment that defers more cases to review can cost more overall than one with a higher per-token price and a better-matched schema.

## 5 A formal core for decision contracts

The propositions of this section apply to any model that honors the contract of Definition 2.1, with state $x$, question $q$ and answer space $\mathcal{Y}_q$ of size $K$, under the premises each one states. Typical premises are calibration or a local head, and none depends on the training objective. Most of them specialize classical results on calibration, selection and cost-sensitive decisions to the contract, while the option-name result is specific to it. The model returns a distribution $\hat{p}$ on $\mathcal{Y}_q$, with top probability $\hat{p}_{\max}$ attained by the answer $\hat{y}$, and a confidence statistic is any function of $\hat{p}$ that a program thresholds. Top-label calibration, which Guo et al. call confidence calibration, requires $\Pr(Y = \hat{y} \mid \hat{p}_{\max} = v) = v$ [46], and classwise calibration requires $\Pr(Y = y \mid \hat{p}(y) = v) = v$ for every answer $y$ [70]. Question isolation means that the distribution for $q_j$ depends on $x$ and $q_j$ alone.

Longer proofs are in Section S1 of the electronic supplement, Table 3 collects the symbols, and Section 8 summarizes benchmark measurements for Proposition 5.3, Proposition 5.4 and Proposition 5.5 [121].

Proposition 5.1 (Isolation and the Fréchet and Hoeffding bounds). *Under question isolation, the returned probabilities $a_1, \dots, a_J$ of binary events $A_1, \dots, A_J$ about one state are marginals, and when each equals $\Pr(A_j \mid x)$ every joint distribution with these marginals is consistent with them. The conjunction is therefore identified only up to the interval*

$$\max\Big(0, \textstyle\sum_{j=1}^{J} a_j - (J-1)\Big) \le \Pr\Big(\bigcap_{j=1}^{J} A_j \;\Big|\; x\Big) \le \min_j a_j, \tag{3}$$

*both ends of which are attained, and a policy that multiplies the marginals asserts an independence that the contract does not supply.*

Proof. The upper bound holds because the conjunction lies in each $A_j$, and the lower bound is Boole's inequality applied to the complements. Nested events attain the upper bound, and events whose complements are disjoint, or cover the sample space when their probabilities sum to at least one, attain the lower bound [1]. □

Under Proposition 5.1, with marginals 0.8 and 0.7, the conjunction can take any value from 0.5 to 0.7, while the product 0.56 holds only under independence. With 5 questions that each return 0.9, the product is 0.59, yet any value from 0.5 to 0.9 is consistent. A program can instead ask one Choice over the joint outcomes, whose size grows exponentially with the number of questions, or state a dependence model.

Proposition 5.2 (Resolution floor). *Let a model report the probability $\hat{p}$ of a fixed event with base rate $\rho$ on a set of values whose smallest positive element is $\delta$, and let $w_0 = \Pr(\hat{p} = 0)$. Every mass-weighted absolute calibration error computed over a partition of the reported values, such as the binned ECE, then satisfies*

$$\mathrm{ECE}(\hat{p}) \geq \left|\mathbb{E}[\hat{p}] - \rho\right| \geq \max\bigl(0, \delta(1 - w_0) - \rho\bigr). \tag{4}$$

*Calibration therefore requires $w_0 \geq 1 - \rho/\delta$, so a model whose reportable values exclude zero cannot be calibrated for an event with $\rho < \delta$.*

Proof. The left inequality is the triangle inequality, since the weighted mean of the differences between mean report and observed rate over the bins is $\mathbb{E}[\hat{p}] - \rho$. The right one holds because every positive report is at least $\delta$, so $\mathbb{E}[\hat{p}] \geq \delta(1 - w_0)$, and calibration sets the error to zero, which gives $\rho \geq \delta(1 - w_0)$ and hence the bound on $w_0$. □

Proposition 5.2 generalizes a bound that Rafe and Das [122] derived for the hosted model. Here resolution means the spacing of reportable values, not the resolution term of the Murphy decomposition [106]. The vendor's API schema states no precision for its probabilities [144], while Rafe and Das [122] measured a two-decimal grid on which exact zero was never returned. With $\delta = 0.01$ and a base rate of 0.002, calibration requires at least 0.8 of the reports to equal zero. A model that places 0.5 of its reports there has a calibration error of at least 0.003, and one that never returns zero has at least 0.008, whatever its ranking quality, because its mean report overstates the prevalence of the event. For open models that return floating-point probabilities the floor is negligible.

Proposition 5.3 (Confidence across answer-set sizes). *Consider the class of confidence statistics $c_K(\hat{p}) = g_K(\hat{p}_{\max})$ with $g_K$ strictly increasing, which contains the normalized maximum*

$$c_K(\hat{p}) = \frac{\hat{p}_{\max} - 1/K}{1 - 1/K}. \tag{5}$$

*A threshold $\tau$ on $c_K$ is the threshold $t_K(\tau) = g_K^{-1}(\tau)$ on the top probability, which equals $\tau + (1 - \tau)/K$ for the normalized maximum. Under top-label calibration within each answer-set size the error rate among accepted answers is at most $1 - t_K(\tau)$, which equals $(1 - \tau)(1 - 1/K)$ for the normalized maximum and increases with $K$.*

Proof. The equivalence of thresholds holds because $g_K$ is strictly increasing. Top-label calibration makes the error rate of accepted answers the mean of $1 - \hat{p}_{\max}$ over answers with $\hat{p}_{\max} \geq t_K(\tau)$, which is at most $1 - t_K(\tau)$. □

Proposition 5.3 shows that one confidence threshold applied to questions of different sizes admits different risks whenever $g_K$ depends on $K$, as it does for the normalized maximum. At $\tau = 0.5$ the normalized maximum accepts a top probability of 0.75 with 2 options, 0.55 with 10 and 0.502 with

255, so the error bound on accepted answers is 0.25 with 2 options and 0.45 with 10. We state the proposition for a class because the vendor reserves the production definition of its confidence value, although a margin or entropy statistic would fall outside it, as Section 4 describes [146]. Every gate should state which quantity it thresholds, and a program that mixes answer-set sizes should set one threshold per question or threshold the top probability itself.

Proposition 5.4 (Cascades, budgets and the three-action rule). *Let a screening stage forward a positive item with event $G$ and a later stage detect it with event $H$. Final recall satisfies*

$$\Pr(G, H \mid Y = 1) = \Pr(G \mid Y = 1)\ \Pr(H \mid G, Y = 1), \tag{6}$$

*and with mean stage costs $c_1$ and $c_2$ the expected cost per item is $c_1 + \Pr(\mathrm{esc})\, c_2 + \Pr(\mathrm{rev})\, c_{\mathrm{R}} + \mathbb{E}[L_{\mathrm{res}}]$, adding screening, escalation, human review and the residual loss of automated decisions. For a calibrated probability $p$ of a positive, flagging is optimal when $p \geq t^* = c_{\mathrm{FP}}/(c_{\mathrm{FP}} + c_{\mathrm{FN}})$. If a review returns the correct label at cost $c_{\mathrm{R}}$, review is optimal exactly on the band*

$$c_{\mathrm{R}}/c_{\mathrm{FN}} \leq p \leq 1 - c_{\mathrm{R}}/c_{\mathrm{FP}}, \tag{7}$$

*which is non-empty when $c_{\mathrm{R}} \leq c_{\mathrm{FP}} c_{\mathrm{FN}}/(c_{\mathrm{FP}} + c_{\mathrm{FN}})$, has positive width when this inequality is strict, and always contains $t^*$. For failure events $D_1, \ldots, D_m$ of a workflow, $\Pr(\bigcup_j D_j) \leq \sum_j \Pr(D_j)$ holds under any dependence.*

A screen that forwards 0.95 of positives and a reader that detects 0.9 of them yield a recall of 0.855, and an evaluation run only on forwarded items reports the second factor alone. With $c_{\mathrm{FP}} = 1$ and $c_{\mathrm{FN}} = 4$ the Bayes threshold is 0.2 [37]. A review cost of 0.5 opens the review band of Chow's rule from 0.125 to 0.5 [17], and the band exists for any review cost up to 0.8. The review budget is the probability mass in this band, and the budget of Rafe and Das [122] is a variant in which a target precision among flagged records sets the upper edge. The union bound is accounting and not an estimate, because each rate it adds must be measured on the inputs that its component receives inside the workflow.

Proposition 5.5 (Option names). *Let the options of a question pair names from a set $\mathcal{N}$ with rubrics from a set $\mathcal{R}$, both of size $K$. For a local head, whose logit for an option depends only on the state, the question and that option's name $n$ and rubric $r$, the logit decomposes uniquely as*

$$\ell(n, r) = \mu + \alpha(r) + \beta(n) + \gamma(n, r), \tag{8}$$

*with a rubric term $\alpha$ centered over rubrics, a name term $\beta$ centered over names, and an interaction $\gamma$ centered over the names for each rubric and over the rubrics for each name. Exchanging the names of two rubrics changes their log-odds by $2[\beta(n) - \beta(n')]$ plus an interaction contrast that vanishes for a binary question and averages to zero over rubric pairs, so swaps with fixed rubrics, averaged over rubric pairs when $K > 2$, identify the name term. Random names that are exchangeable and independent of the rubrics give expected log-odds that depend on the rubrics alone, so the name term becomes constant across options in expectation. The head is name-invariant when its distribution over rubrics is the same under every assignment of names. For $K > 2$ this holds if and only if $\beta \equiv 0$ and $\gamma \equiv 0$, and for a binary question if and only if $\beta \equiv 0$, with $\gamma$ then unidentified and without effect.*

Proposition 5.5 separates the identifier that a program uses for an option from the rubric that defines it. Take a binary question with rubric contrast $\Delta\alpha = 1.0$ toward the positive rubric and name contrast $\Delta\beta = 1.5$ toward the name yes. The aligned naming gives log-odds 2.5 and probability 0.92 on the positive rubric, while the swapped naming gives log-odds $-0.5$ and probability 0.38, so the decision follows the name. Half the difference of the two log-odds recovers $\Delta\beta$, half their sum recovers $\Delta\alpha$. Random strings without an average name effect give expected log-odds 1.0, which correspond to a probability of 0.73. Sun et al. [139] applied this design to the hosted model and

to open-weight models and reported that exchanging polar names changed many more decisions than exchanging neutral names, while random character strings restored the neutral regime. The design rule is to use neutral identifiers, place the meaning in the rubric, and test every deployed schema with the swap.

Proposition 5.6 (Ordinal outputs). *Let a Score question have ordered levels* $1, \dots, L$ *with* $L > 2$*, probabilities* $\hat{p}_k$*, cumulative probabilities* $\hat{F}_k = \sum_{j \le k} \hat{p}_j$ *and expected level* $m(\hat{p}) = \sum_k k\, \hat{p}_k$*. The expected level does not determine the cumulative probabilities, and a forecast can satisfy* $\mathbb{E}[Y \mid m(\hat{p})] = m(\hat{p})$ *while its cumulative probabilities are miscalibrated. The ranked probability score*

$$\mathrm{RPS}(\hat{p}, y) = \frac{1}{L-1} \sum_{k=1}^{L-1} \big(\hat{F}_k - \mathbf{1}\{y \le k\}\big)^2 \tag{9}$$

*is strictly proper, and ordinal calibration, defined as* $\Pr(Y \le k \mid \hat{F}_k = v) = v$ *for every* $k < L$ *and every* $v$*, makes a Bayes threshold on the level optimal for every cost ratio.*

Proposition 5.6 warns against consuming the expected level of a Score, the probability-weighted mean that Section 4 describes. On 5 levels, a forecast with 0.8 on the middle level and 0.1 on each neighbor and a forecast with 0.5 on each extreme share the expected level 3.0, yet they give 0.1 and 0.5 to a level of at least 4. The squared error of the expected level cannot separate them. The RPS of Epstein [38], where lower is better, gives 0.005 and 0.25 for outcome level 3 and 0.455 and 0.25 for level 5. Applications should consume the cumulative probabilities that their thresholds need and evaluate Score outputs with the RPS or with Brier scores on threshold events [11, 45].

## 6 The ecosystem census

The ecosystem census enumerates every publicly recorded artifact that instantiates the decision contract of Definition 2.1 or that operates one, screened by the method of Section 2. At the first cutoff it separated 101 original implementations, grouped into 67 recipe families distinguishable by author, backbone and training recipe, from 259 derivatives that reused a T1 checkpoint without an independent recipe of their own. Most derivatives are format conversions for a different runtime, at 142 of the total, while smaller groups merge checkpoints, quantize them, or fine tune them for a new domain or language, at 4, 9, 9 and 10 entries respectively. The remaining 85 derivatives fit none of these groups. A further 54 repositories are datasets built to train or evaluate a decision model. Section S5 of the supplement lists every original implementation individually. The second sweep found 2,645 further Hugging Face entries created by 28 October 2026. We read by hand every card that a pattern rule had tiered as an original implementation, confirmed 278 originals in 149 recipe families, some of them new checkpoints of families above, and moved 42 to derivatives or out of scope. GitHub search found 15,788 further repositories that pass the relevance screen.

Table S6 in the electronic supplement condenses the census to the twenty families with an independent card or vendor page, five of them released after the first cutoff. It records each family's backbone, decision head, training objective, typed primitives and serving temperature or calibration as shipped, verified against its own card or repository. The serving temperature that each family fits before returning a probability differs widely across families, as the final column of Table S6 records. Most open implementations attach a low-rank adapter or a compact classification head to a public backbone rather than train a new base model from scratch. Kev, Bespoke-Nimble and the pngwn scorer each add a small trained component to a public Qwen checkpoint this way [8, 115, 116], while Laya and its multilingual sibling instead train an encoder classification head on a public ModernBERT or mmBERT backbone [24, 25]. The recorded release dates in Table S7 span

2026-09-16 to 2026-09-24, consistent with a shared reaction to one launched interface rather than independent convergence on it.

The census datasets include benchmarks built against exactly this contract. LocalLLaMA's typed-decisions collection is the most prominent one [90]. Its own card reports a teacher self-agreement ceiling of 0.735, against which every model's accuracy on it should be read. Luni's laya-jev-benchmark is a third-party comparison set assembled after the launch [91]. Two tooling patterns recur among the harness repositories the census records separately from the models. Palmer's kev bundles a local server that reimplements the hosted vendor's request and response shape, so an existing client software development kit can point at an open checkpoint instead of the hosted service [114]. A separate community harness instead points that same client at the hosted Jev itself, through an early decisions endpoint OpenRouter added after the launch [109]. Among the screened repositories, the census counts 48 local servers and 25 evaluation packages.

The hosted side of the class widened after the launch, and five providers served or released a typed decision model by the cutoff. TypeSafe released jev-1.14.0 on 14 October 2026 at an unchanged price and context, moved both public aliases to it, and added Box, a fourth question type that returns a single bounding box in an image [145, 150]. Microsoft-Decision-1 is a post-trained Qwen3.5-9B served through Microsoft Foundry and OpenRouter with the same three primitives, and its documentation asks callers to read Score outputs as an ordering rather than as calibrated ratings [99, 100]. MachineFi's Trio-Spark serves Choice questions over at most 8 options and reports an expected calibration error of 0.019 on its own panel and adds that the value need not hold in another application [95]. Cloudflare's Clef and Perplexity's pplx-decider ship open weights under the Apache license, and Clef is also hosted [20, 118].

## 7 Evidence synthesis

The evidence ledger holds 248 graded studies, which form 247 independent clusters because two share an author group and a testbed [78, 79]. Each study was graded on the five risk-of-bias domains of Section S3, and 3 reached low overall, two of them the present authors' own, 103 some concerns and 142 high. High ratings number 45 for reference-label independence, 10 for sampling and reporting, 65 for version pinning and run dates, 55 for tuning-budget parity and 93 for test-set exposure. Exposure is therefore the most frequent threat. The largest group of evaluations scores public benchmarks that predate the models without a contamination check, and reading those ratings as unclear would leave 61 studies graded high.

Each study enters a summary as one value, the median of its rows, and each summary reports the median over clusters with a percentile bootstrap interval when at least three clusters share the estimand. Three regularities organize the synthesis. The contract travels, and robustness to option names belongs to each checkpoint, so it must be tested per model. Calibration is conditional, because it belongs to the model and the task together and is audited per schema. Escalation earns the saving, because a confidence that ranks errors lets a gated policy send only uncertain cases to the stronger model, and its value depends on the price gap and on routing accuracy.

### 7.1 The contract travels

Open label-conditioned heads on public backbones reproduced the typed contract within days of the hosted release, as Section 6 records. Where an open model and the hosted model were scored on the same items, the median gap across 63 clusters, best open minus hosted, was 1.3 points, with a bootstrap interval from −0.7 to 3.2. The best open model was ahead in 37 clusters and not ahead in 26. An exploratory split by how the open model was built, which we coded after the first synthesis of this version and recoded without the computed contrasts (Cohen's kappa 0.86), separates two effects. For open decision models used zero-shot, the median gap, open minus hosted, was −0.7

points over 35 clusters, with an interval from −4.0 to 2.2. For open models trained or selected on the target task it was 3.9 points over 29 clusters, with an interval from 0.9 to 6.1.

Every task-trained contrast comes from a study rated unclear or high for tuning parity, usually because its authors built the model they report, so this ledger cannot separate training on the task from self-report. The 9 parity-controlled clusters all compare zero-shot models, with a median gap of −0.4 points [for example 28]. Matched contrasts of a typed readout with answer fine-tuning of the same backbone on the same data attribute no more than a few points to the typed head itself [86, 126, 166]. The contract therefore transfers to open models, and an accuracy lead over the hosted model appears only for models trained on the task, all reported by their own builders.

Robustness to the presentation of options varies by model. Under an option-name polarity swap, an open marker head that led the hosted model with aligned names moved to a gap of −34.9 AUROC points [139]. One seeded shuffle of option order changed 4.7, 7.6 and 9.1 percent of the answers of three decision models on the same items, while a head trained for order invariance changed none [142]. Section 8 finds under one harness that option-name sensitivity varies by checkpoint, even within one recipe family of Section 6.

Over every coded reading of each comparator class, the median gap to frontier language models, hosted minus comparator, was −3.7 points over 29 clusters, with an interval from −7.8 to −1.6. Against smaller or open language models the median was −1.0 over 31 clusters, and against models trained on the task it was −7.9 over 25. The hosted model was cheaper by a median factor of 37 over 43 clusters and faster by 5.3 over 42.

Table S11 in the supplement sets each vendor launch claim beside its independent status, with the vendor's own qualifiers. A claim is supported when every independent record agrees with it, and supported in direction when all or all but one record agree on its direction but not on its stated magnitude or scope. It is not supported as stated when the records point both ways on the property itself, and definitional when it holds by construction of the output type. Developer self-reports are shown but do not count as independent.

### 7.2 Calibration is conditional

The hosted model's raw calibration error varied by a factor of 77 across 47 clusters (Figure 4a). It ranged from 0.007 in a study of prevalence shift [60] to 0.541 on an attack-detection cohort where most inputs were attacks [74]. The present authors' document-coding study used an exact estimator on the probability grid and still rejected calibration, with a Spiegelhalter statistic of −5.3 and a frontier arm better calibrated at 0.0027 [122]. Binning does not explain the spread. Alternative binnings in the parity-controlled study moved the error 88 times less than the between-study spread. Under one fixed scheme of ten equal-width bins, the error still ranged from 0.027 to 0.541 over 15 clusters. The estimand contributes, because a median over tasks exceeded the pooled error on the same decisions by 0.121 in one study [47]. Within the parity-controlled study, where the estimator is fixed, the error still ranged from 0.063 to 0.538 across tasks [58].

Recalibration depends on the data used to fit the map (Figure 4b). Across 21 studies that report both values, the recalibrated error ranged from 0.024 to 1.9 times the raw error, so one temperature fit raised it [15]. In-domain fits ranged from 0.024 to 0.89 of the raw error over 9 records, and transferred fits from 0.36 to 0.96 over 7. A temperature that reduced one open model's error on one evaluation set left it unchanged on that model's holdout [9]. In one study, one shared temperature moved the count of language models below the hosted model's calibration error from 3 to 15 of 19 [58]. The native advantage over recalibrated language models is therefore fragile.

The ledger also documents regions where confidence and correctness part ways. On an empathy task, the hosted model's accuracy at high confidence was barely above the base rate [58]. An open English model answered every item in one non-Latin script incorrectly at high confidence

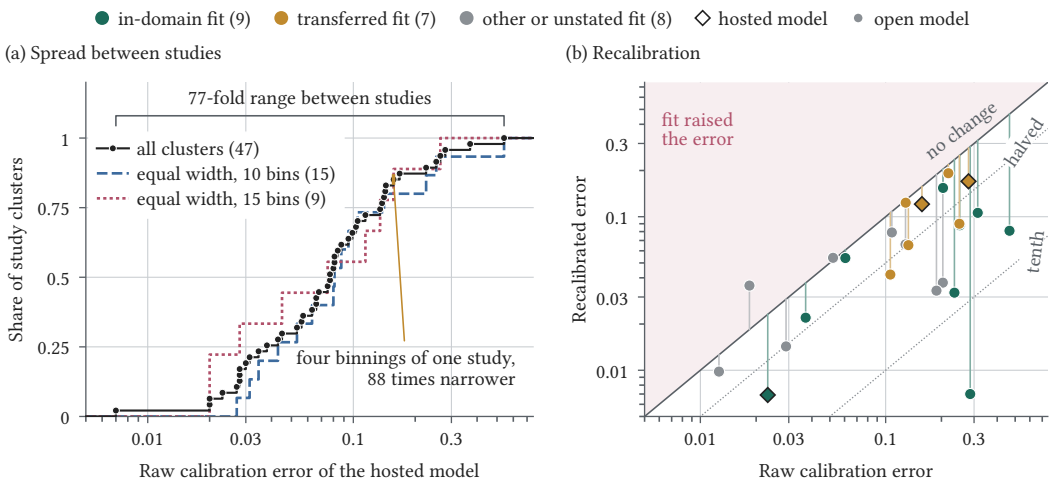



**Fig. 4.** Calibration error in the evidence ledger. (a) Distribution of the hosted model's raw error over study clusters, overall and under two fixed binning schemes, with the spread of one study over four binnings. (b) Raw against recalibrated error for every estimate reporting both, by the data used to fit the map.

[25]. Confidence still orders errors in aggregate, with accuracy rising sharply from the low- to the high-confidence band [81]. This spread recurs across primitives and studies, so a program should audit each schema's calibration on its own inputs rather than assume one global value.

### 7.3 Escalation earns the saving

Figure 5 plots retained quality against cost fraction and against the escalated share for every cascade record that reports them. Across 23 studies, confidence-gated escalation to a stronger model kept a median 0.99 of that model's quality, ranging from 0.67 to 1.2, where a value above one means the cascade beat the stronger model alone. It paid a median 0.56 of the stronger model's fee over 15 studies and escalated a median 0.34 of decisions over 16. These studies define cost differently, as a fallback fee, a task-cost fraction or a per-episode cost, and fix thresholds on pilot data, on a selection split or without saying when. The confidence ranking carries part of the saving, since one study's gated escalation reached a risk-coverage area of 0.019 against 0.145 for random escalation at matched coverage [162].

The saving shrank as the price gap to the partner model narrowed, and with a low-priced open partner the cascade cost most of what the partner alone cost, and more than the partner alone at a stricter threshold [58]. Where routing was already accurate, a cheap generative cascade that escalated on its own judgment matched the decision-model cascade at 0.72 of its cost [162]. Two designs stay outside these estimates. A hybrid sends a fixed sparse set of states to a flagship model [80], and a review budget set by a precision target is a deferral budget for human review rather than a cascade [122].

The literature at the cutoff leaves several questions open. Only 9 accuracy comparisons of an open model with the hosted model hold tuning parity, and none of them involves a model trained on the task. Cascade studies still use no common cost definition. Several studies of the second sweep evaluate jev-1.14.0 [13, 28, 50], but none follows one deployment across the alias change of Section 4. Section 8 adds, outside the ledger, a swap test of every decision model and cascades under one cost definition that escalate to the hosted decision model rather than to a frontier model.

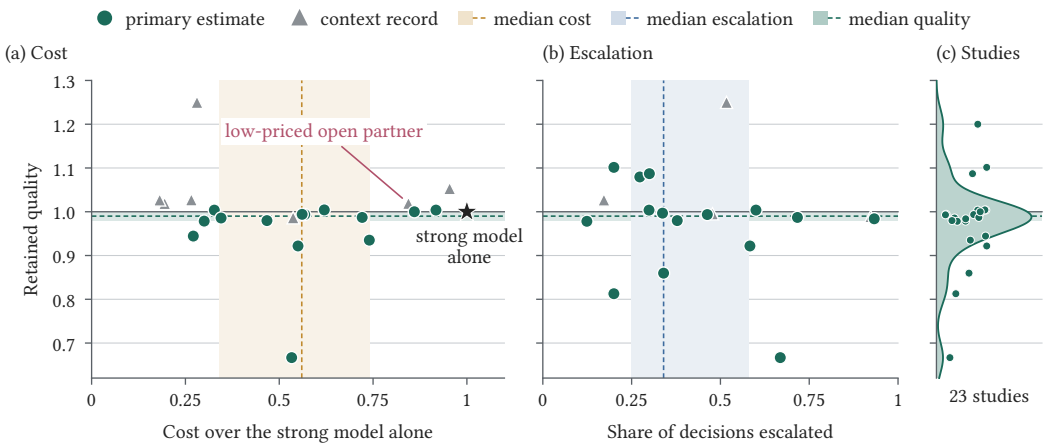



**Fig. 5.** Confidence-gated cascades in the evidence ledger, each value a ratio to the strong model alone on the same work as the study defines it. Retained quality against (a) cost and (b) the share of decisions escalated, and (c) its distribution over studies. Dashed lines mark the medians over studies of cost, escalated share and retained quality, and shaded bands of the same color their 95% bootstrap intervals.

Section 9 states the cost identity of these cascades, while their retained quality stays an empirical question.

## 8 A same-harness measurement

The companion benchmark measures the hosted model and several open decision models under one harness [121]. Its datasets are typed-decisions [90], CLINC-150, 4 computational social science tasks, GoEmotions and Banking77. As the authors' own measurements, these results are never pooled with the ledger. Every model receives the same semantic request, with the same state, question and option texts in the same order. Each open checkpoint runs at a pinned revision, and the hosted model runs at jev-1.13.0 with 3 repeats. The rendering into tokens stays each model's own, and it matters. The this-that-model-1.0 adapter joins each option key to its description, and a rerun with descriptions alone raised its CLINC-150 accuracy by 6.8 points, with an interval from 4.5 to 9.3. The open decision models returned identical answers on every retest. The companion paper reports four open generative models from three developers as comparators.

Table 4 gives the headline results, with accuracy scored against the teacher-built labels of typed-decisions. The hosted model scored 0.732, beside the teacher self-agreement reference of 0.735. The nearest open decision model, Kev-9B, differed from it by −1.9 points, with a state-clustered interval from −4.0 to 0.1 and a Holm-adjusted p-value of 0.37. The gap is therefore not distinguishable from zero.

Conventional baselines place these results by the labels available. Classifiers trained on the CLINC-150 training split reached 0.940 and 0.948 with every intent offered, above every decision model, while a zero-shot entailment model reached 0.507. On typed-decisions, soft-label regressions on sentence embeddings trained with the teacher's labels reached 0.702, while a fine-tuned DeBERTa-v3-large reached 0.793, above the hosted model. Used without task labels in this harness, the hosted model, Kev-9B and Nimble-9B therefore far exceeded zero-shot entailment on intents, while small classifiers trained on the task's own labels were more accurate still. The confidence of the hosted model also leaves out-of-scope requests unguarded. At a held-out threshold for 5 percent in-scope

**Table 4.** Headline same-harness results of the companion benchmark [121], on typed-decisions (D1) with neutral identifiers except CLINC@150, the CLINC-150 accuracy with all intents offered. Bold marks the best value in each column, the hosted model exposes no raw output, and decider-2b and this-that-model-1.0 were trained on CLINC-150. Cost is per 1,000 decisions, with a request's cost shared among its questions, and latency is per request, with open models run one request at a time.

| Model | Accuracy | Calibration (ECE) | | Robustness | | Cost | |
|---|---|---|---|---|---|---|---|
| | D1 | raw | T-scaled | flips/100 | CLINC@150 | USD/1k | $p_{50}$ ms |
| Hosted, jev-1.13.0, API at 0.042 USD per million tokens | | | | | | | |
| Jev 1.13 | **0.732** | – | 0.032 | 50.5 | **0.913** | 0.0079 | 138 |
| Open, pinned revisions, Colab L4 at 0.48 USD per hour | | | | | | | |
| Laya (EN) | 0.366 | 0.232 | 0.068 | 91.3 | refused | 0.0016 | 60 |
| Laya (ML) | 0.349 | 0.279 | 0.044 | 86.8 | 0.005 | **0.0009** | **32** |
| Kev-0.8B | 0.451 | 0.161 | 0.070 | 14.3 | 0.685 | 0.0025 | 84 |
| Kev-9B | 0.712 | 0.092 | 0.047 | 6.5 | 0.850 | 0.0147 | 590 |
| decider-2b | 0.588 | 0.139 | 0.036 | 4.8 | 0.860 | 0.0109 | 426 |
| this-that-model-1.0 | 0.597 | 0.196 | 0.040 | 4.3 | 0.863 | 0.0020 | 71 |
| Nimble-9B | 0.685 | **0.057** | **0.026** | **4.2** | 0.860 | 0.0532 | 2056 |

risk, it still accepted 0.310 of out-of-scope CLINC-150 requests, with an interval from 0.142 to 0.502, at an in-scope coverage of 0.92. Kev-9B accepted 0.015 at a coverage of 0.61.

Figure 6 shows the option-name swap in panel (a) and the answer-set sweep in panel (b). Swapping the no and yes keys against a fixed rubric changed 50.5 to 97.2 of every hundred hosted-model answers, far above the 0.4 to 1.3 answers per hundred that changed across repeats with aligned names. The two Laya checkpoints changed 71.4 to 98.8, while the Kev, Nimble, decider and this-that checkpoints changed 0.2 to 14.3. Sensitivity to names therefore varies by checkpoint, even within one recipe family, and each deployment must run the renaming test of Proposition 5.5.

Panel (b) varies the answer-set size of Proposition 5.3, and the hosted model's calibration error reached 0.032 at 150 options. Two further permutations of the distractors moved accuracy by at most 0.013 for the hosted model and 0.035 for the open models, except Laya (EN) at 20 options with 0.058. Both Laya checkpoints held 0.830 or more at 20 options, fell to 0.065 and 0.020 at 50, and at 150 the English head refused the request, as the cardinality row of Table 6 records. A replay of the English head's own tokenizer traces both failures to its option budget. At 50 options the option texts needed 302 tokens against a head budget of 192, so every option was cut to its key, and at 150 options 24 options fell beyond the input limit.

Calibration as shipped also depends on the model. Raw calibration error on typed-decisions ranged from 0.057 for Nimble-9B to 0.279 for Laya (ML), and the hosted model, which exposes no raw output, reached 0.033 as shipped. One temperature per primitive type, fitted on a disjoint slice of the same dataset, placed Nimble-9B at 0.026, below the hosted model's 0.032, which fits the second regularity. Their intervals still overlap, from 0.020 to 0.058 and from 0.023 to 0.055.

The cascades of Proposition 5.4 price every decision from its own recorded API tokens at list price or its GPU time on a Colab L4 at 0.48 USD per hour. Every threshold is chosen out of fold, on the calibration slice for typed-decisions and by item-disjoint cross-fitting elsewhere, and the intervals refit it in every bootstrap replicate. On typed-decisions, where Kev-9B and Nimble-9B already cost more per decision than the hosted API (Table 4), escalating from them matched the hosted model at 2.37 and 7.55 times its cost. On CLINC-150 neither exceeded the hosted model's accuracy, at 1.08 and 2.33 times its cost, with Kev-9B escalating 0.56 of its decisions. Of the smaller first stages, decider-2b and this-that-model-1.0 saved cost there without a significant accuracy gain,

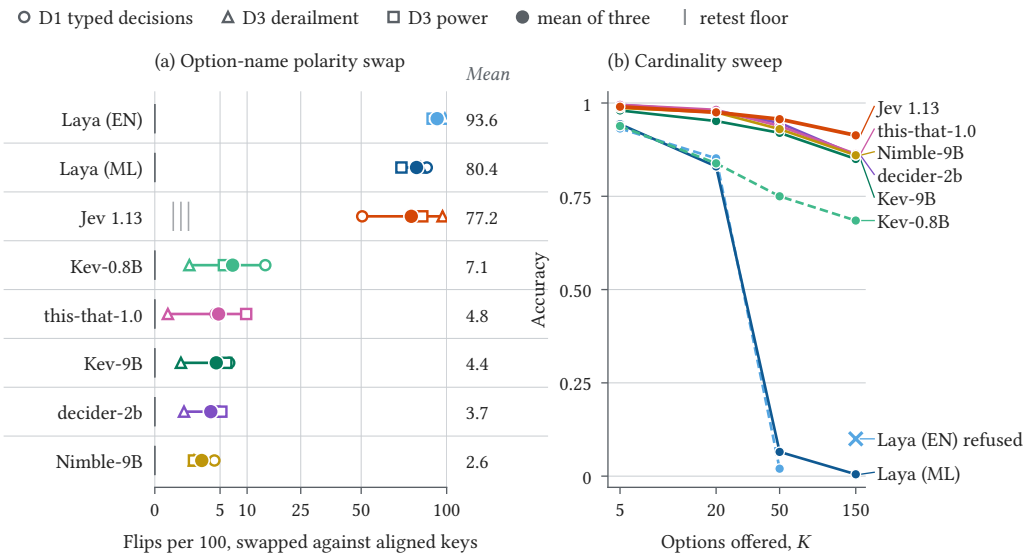



**Fig. 6.** Option-name swap and answer-set size for the decision models of the companion benchmark [121], whose generative comparators are reported there. Panel (a) gives flips per hundred decisions when the no and yes keys are swapped, and panel (b) gives CLINC-150 accuracy as the options grow.

while Kev-0.8B and Laya (ML) cost more than the hosted model. The this-that-model-1.0 cascade reached 0.927 at 0.43 of the hosted cost, a gain of 1.3 points with an interval from −0.7 to 2.8. The decider-2b cascade gained 1.2 points, from −1.2 to 2.0, at 0.58 of the hosted cost. Both first stages were trained on the CLINC-150 training split, so these results are in-distribution, and both cascades match rather than exceed the hosted model. A 9B first stage thus cost more than the hosted model alone on both datasets. The cheaper small first stages bought savings rather than accuracy, as the price-gap and routing-accuracy clauses of the third regularity predict.

## 9 Patterns and a failure atlas

A pattern in this survey is a recurring way of composing decision-model calls with ordinary code, and each pattern consumes a probability quantity whose meaning Section 5 fixes. The vendor documents its own architectural patterns, namely speculative fan-out, confidence-gated routing, composite scoring and intent routing, and it claims a cost, speed, reliability or safety benefit for each [151]. A claimed benefit is not a guarantee, so the guarantee status in Table 5 comes from the propositions or the ledger, and Section S6 of the electronic supplement gives every entry in full.

Several patterns rest on a proposition and inherit its assumptions. The gate and the review budget follow Proposition 5.3 and Proposition 5.4, which bound the error among accepted answers and place review where it is optimal. Both hold only under calibration, which belongs to the model and the task together, so these guarantees are audited per schema. In Rafe and Das [122], the review budget is a precision-targeted deferral budget, not a cascade result. Marginal-to-joint reconciliation follows Proposition 5.1, so a program asks one Choice over the joint outcomes or carries the interval instead of a product. Option-name neutralization follows Proposition 5.5, because exchangeable random names remove the name term in expectation. Paraphrase self-consistency therefore has a proof when it varies names and none when it varies wording.

**Table 5.** Condensed pattern catalog, grouped by role. The guarantee is ● proved under the calibration its proposition assumes, ◑ empirical or conformal, or ○ none, and rows name the failure modes of Table 6 it mitigates.

| Pattern | Quantity consumed | Guarantee | | Rows | Evidence |
|---|---|---|---|---|---|
| Threshold and route | | | | | |
| Gate | Top, positive or cumulative probability against a threshold | ● | Under calibration, Propositions 5.3 and 5.4; conformal in principle | 10 | Ibrahim and Zaki [58]; Li et al. [81] |
| Screen-then-read | Screen probability, then read-stage distributions | ◑ | Recall identity proved, Proposition 5.4; factors empirical | 5 | Rafe and Das [122] |
| Candidate-then-select | Choice distribution over a short candidate list | ◑ | Conformal if the list is a conformal set; otherwise empirical | 9 | Mapika [97]; Wu and Lim [162] |
| Hierarchical fallback | Chained Choice distributions | ◑ | Stage product proved, Proposition 5.4; accuracy empirical | 9 | TypeSafe AI [151]; Rafe and Das [121] |
| Escalate or defer | | | | | |
| Confidence-gated cascade | Top probability against an escalation threshold | ◑ | Cost identity proved, Proposition 5.4; quality empirical | 2, 6 | Li et al. [81]; Wu and Lim [162] |
| Judge-and-escalate | Probability of the preferred answer | ◑ | Empirical | 2, 6 | Li et al. [81]; Huang et al. [56] |
| Review budget | Calibrated probability in the review band | ● | Under calibration, Proposition 5.4 | 14 | Rafe and Das [122] |
| Combine answers | | | | | |
| Speculative fan-out | Marginals of many isolated questions | ○ | None beyond each marginal, Proposition 5.1 | – | TypeSafe AI [151]; Balian, Rob [4] |
| Composite scoring | Several marginals combined in code | ○ | None, Propositions 5.1 and 5.6 | 7 | TypeSafe AI [151]; Robitza [129] |
| Marginal-to-joint reconciliation | Marginals of events about one state | ● | Proposition 5.1 | 7 | TypeSafe AI [149] |
| Paraphrase self-consistency | Spread across wordings, orders and names | ● | For random names, Proposition 5.5; else empirical | 7, 8 | Sun et al. [139]; Ren et al. [126] |
| Harden the input and the contract | | | | | |
| Deterministic pre-parsing | None for the parsed part | ◑ | Exact for the parsed part; else empirical | 2, 3, 4, 12 | Palmer, Jared [116]; TypeSafe AI [149] |
| Versioned contract | None; pins version, schema, names, temperature | ○ | None; procedural | 8, 14 | Bespoke Labs [8]; Palmer, Jared [115] |

The remaining patterns carry empirical support or none. For the confidence-gated cascade and judge-and-escalate, Proposition 5.4 supplies the cost identity, while the retained quality is measured in 23 studies with different cost definitions [58, 81, 162]. The saving comes from escalation gated by a confidence that ranks errors, and its size depends on the price gap and on routing accuracy. Speculative fan-out and composite scoring have no guarantee beyond the calibration of each marginal. Candidate-then-select and the gate admit a conformal guarantee in principle [2, 3], which no ledger study applies to a decision model.

The versioned contract has no formal guarantee, and the census motivates it most directly. Bespoke-Nimble-9B changed revision behind one repository name, from a checkpoint served at a fitted temperature to one shipped at its default temperature with a larger answer space and a warning not to reuse the earlier fit [8]. Kev-0.8B also moved its serving temperature between

**Table 6.** Failure atlas of decision models. Each cell has three slots, filled when the vendor or developer documents the failure mode (●V), a ledger study measures it (●L) or the authors' companion benchmark, which is not risk-of-bias graded, measures it (●B).

| Failure mode | Hosted | Open decision models | | | | | Generative |
|---|---|---|---|---|---|---|---|
| | Jev | Laya | Kev | Nimble | this-that | decider | Gen. comp. |
| VENDOR JAGGEDNESS MODES | | | | | | | |
| 1 Negation and scoping | ●●○ | ○●○ | ○○○ | ○○○ | ●○○ | ●○○ | ○●○ |
| 2 Counting and arithmetic | ●●○ | ○●○ | ○○○ | ○○○ | ○●○ | ○○○ | ○○○ |
| 3 Dates | ●●○ | ○○○ | ●●○ | ○○○ | ○○○ | ○○○ | ○○○ |
| 4 Indirection | ●○○ | ○○○ | ○○○ | ○○○ | ○○○ | ●○○ | ○○○ |
| 5 Distractor state | ●○○ | ○●○ | ○●○ | ○○○ | ○○○ | ○○○ | ○○○ |
| 6 Adversarial state | ●●○ | ○●○ | ○○○ | ○○○ | ○○○ | ○●○ | ○●○ |
| 7 Cross-phrasing inconsistency | ●●○ | ○●○ | ○●○ | ○●○ | ○●○ | ●●○ | ○●○ |
| FURTHER FAILURE MODES | | | | | | | |
| 8 Option-name polarity | ○●● | ●○● | ○●● | ○○● | ○○● | ●○● | ○○● |
| 9 Cardinality collapse | ○●● | ●●● | ○●● | ○○● | ○○● | ●○● | ○○● |
| 10 Ordinal numerics | ○●○ | ○○○ | ○●○ | ●○○ | ○○○ | ○●○ | ○○○ |
| 11 Language transfer | ○●○ | ●●○ | ○●○ | ○●○ | ○○○ | ●○○ | ○●○ |
| 12 Physical magnitudes | ○○○ | ○○○ | ○○○ | ○○○ | ○○○ | ○○○ | ○○○ |
| 13 Top-confidence inversion | ○●○ | ●●○ | ○●○ | ○○○ | ○○○ | ○●○ | ○○○ |
| 14 Shipped over-confidence | ○●● | ●●● | ○●● | ○●● | ○●● | ○○● | ○●● |
| EVALUATION | | | | | | | |
| 15 Teacher-consensus benchmark bias | ○●○ | ○●○ | ○●○ | ●○○ | ○○○ | ○○○ | ○○○ |

checkpoints under an unchanged name [115]. Pinning the model version, the schema, the option names and the temperature, and re-auditing calibration after any change, makes such a change detectable [122].

Table 6 crosses failure modes with the hosted model, the open families of the companion benchmark (Kev pooled over its two checkpoints) and a generative column that pools the four open generative comparators of that benchmark. A filled slot means a source documents or measures the failure mode for that model. It does not give the direction, since several ledger rows are favorable, such as the false-negation test of Huang et al. [56] and the order perturbation of Zhang et al. [170]. Rows on jaggedness follow the vendor's own failure list [149], the final row is an evaluation failure, and a script sets the benchmark codes from the companion results [121].

The evidence is uneven across failure modes and across models. For the hosted model, indirection and distractor state rest on the vendor page alone, while counting, option order, option-name polarity and cardinality collapse are measured independently. A polar-name swap flipped 32.50 answers per hundred [139], and accuracy changed by −6.7 points as the option count grew [162]. For the open families, many measured cells come from the developer's own card or paper, so a vendor code and a ledger code in one cell can record one source. Examples include the Khmer result of Laya, at 0.000 accuracy with 0.952 confidence [25], and the multi-step arithmetic of this-that-model at 0.560 accuracy [16]. In independent tests of option order from the second sweep, native Laya flipped 30.6 percent of decisions under one shuffle against 5.3 for the hosted model [55]. The generative comparator has the fewest vendor and ledger codes, which reflects where studies looked. No model has evidence on physical magnitudes in the ledger, the cards or the vendor pages.

Read together, the tables pair each failure with the patterns designed for it. For counting, dates and indirection, the vendor's remedies amount to deterministic pre-parsing, which the Kev card supports for dates [116]. Hierarchical fallback or candidate-then-select addresses cardinality collapse, and the decider card supports it with sampled label sets [97]. Neutral names and the swap test address

option-name polarity, and in-domain recalibration under a versioned contract addresses shipped over-confidence. Top-confidence inversion, and the confident errors of language transfer, defeat every pattern that thresholds confidence, so for them the catalog offers only an audit of each schema on the inputs it will receive.

## 10 Applications by evidence grade

Table S10 in the supplement lists every ledger study by application area, reference-label type and risk-of-bias grade, and this section reads the areas by the strength of their references. Document coding at population scale rests on human-adjudicated labels, with crash narratives scored against trained coders [122] and contract clauses against expert annotations [170]. Annotation for science turns a decision model into a labeling instrument for social-science coding [58], scientific literature [29], video quality [129] and a population kernel [80], with references that range from human-coded corpora to a computed metric. Judging grades another system's output against a strong language-model judge [81] or against radiologists' consensus error counts [56].

The remaining areas rest mostly on labels that a benchmark, an automated judge or the system itself supplies. They cover agent control and memory [34, 61, 154, 162], guardrails, routing and edge orchestration [47, 78, 79, 124, 126], games and real-time control [16, 94], and vision [166]. Most of these studies were rated high risk of bias, so their accuracy values rest on weaker references than those of document coding and annotation.

## 11 An audit protocol and reporting standard

An evaluation of a decision model is reproducible only when the artifacts that generate its predictions are pinned as tightly as the schema that consumes them. We extend the reproducibility record of the companion benchmark paper into a decision contract record, which lists what a decision-model evaluation must fix before it runs and disclose once reported [121]. The record holds the provider or checkpoint with its resolved revision and the date that revision was confirmed. It holds the serving temperature and any decoding parameter in force, rather than a documented default, and the literal schema sent to the model, including the option identifiers and the rubric bound to each one. It also holds the smallest positive reportable probability and the share of reports equal to zero, the state's serialization and truncation rule, and the hardware or provider behind any latency or cost value. Two rows of the evidence ledger trace their disagreement to exactly one of these elements moving between access dates, so the record is built to make such a move visible.

A reporting checklist turns the record into statements a reviewer can check, following the model that TRIPOD and TRIPOD+AI set for a prediction model and CONSORT-AI and SPIRIT-AI set for a trial that embeds one [22, 23, 88, 127]. Broad benchmark suites standardize scenarios and prompts across language models [82], whereas the record fixes the artifacts of one decision contract that a deployment consumes. PROBAST and QUADAS-2, together with the QUADAS-AI and STARD-AI protocol for diagnostic models, supply the domains an evaluation should rate rather than assume [103, 134, 158, 161]. PRISMA 2020 sets a precedent for stating how a body of evidence was searched and screened [112]. The model card and the datasheet describe a component before deployment, and the audit framework of Raji and colleagues audits one afterward [42, 102, 123]. The literature on leakage and reproducibility motivates the version-pinning and exposure items below [68, 119].

The items answer patterns that recur across the studies surveyed rather than any single one. A served checkpoint can change behind an unaltered name, a temperature one release documents can differ from the value another ships, and a confidence value can be reported without its formula. A label from a teacher model or a synthetic procedure can sit beside a label a person adjudicated, and a speed or cost claim can appear without its hardware, concurrency, or workload. Table S5 in the electronic supplement states, for each reporting item, the quantity to report and the proposition,

failure mode or other source that motivates it. Section S4 of the electronic supplement explains every item of the full checklist and gives the card template that the checklist complements.

**Reporting in brief**

Pin the model and its revision, the serving temperature, the schema with its option identifiers and rubrics, the state encoding, and the access dates of every hosted service. Sample with the provenance of the reference labels stated, together with the sample size and any exposure of the test items in public data. Report the calibration quantity that fits each primitive with its binning, and name the quantity that every gate thresholds. State a cascade result as retained quality against a cost fraction under one cost definition, and state a review budget as a band of probabilities. Treat a changed revision as a new predictor, and supersede earlier values instead of overwriting them.

## 12 Open problems and a research agenda

The evidence synthesized in Section 7 leaves six problems open, and each can be settled by a study whose design and passing condition are fixed before it runs. For each problem, Table 7 gives a study design and an evidence-of-progress criterion that a future study can pass or fail, both built on the propositions of Section 5 and the reporting standard of Section 11.

**Table 7.** A research agenda for decision models, with a study design and a pass-or-fail criterion for each open problem. The line under each problem names the result it builds on.

| Problem | Study design | Evidence-of-progress criterion |
|---|---|---|
| Schema transfer of calibration<br>Section 7.2 | Fit a temperature or recalibration map on some schema families and hold out whole families. Compare the transferred fit with an in-domain fit and with no fit. | On every held-out family, the transferred fit's calibration error lies inside the bootstrap interval of the in-domain fit, under a stated binning and at matched accuracy. |
| Joint uncertainty under isolation<br>Proposition 5.1 | Pose conjunction, implication and exclusion events about one state. Compare a dependence-aware output with the product of marginals and with one Choice over the joint outcomes. | Joint probabilities lie inside the Fréchet and Hoeffding interval and score a lower Brier score on held-out joint events than the product of marginals. The number of calls grows at most linearly with the number of questions. |
| Evidence-aware abstention<br>Proposition 5.4 | Pair each state with a copy from which adjudicated supporting evidence is removed. Compare abstention on an evidence signal with abstention on the top probability. | At matched coverage, the evidence signal gives lower risk on the evidence-removed states than the top probability does, without lower accuracy on the complete states. |
| Contract versioning<br>Section 11 | Re-run a frozen, version-pinned panel of schemas, name swaps and calibration items on consecutive releases of one model. A same-version rerun is the control. | Flip rates and calibration change between releases are reported with intervals, and every change beyond the same-version rerun noise coincides with a change in the stated contract version. |
| Multimodal state<br>Proposition 5.5 | Present the same items with the state as text and as an image. Repeat the option-name swap and the calibration audit on both forms. | The image-state calibration error and swap flip rate lie inside the intervals of the text-state run, including at option counts beyond the training range. |
| Benchmarks without teacher-consensus bias<br>Table 6, row 15 | Label a stratified subset by human adjudication with measured agreement between annotators. Score the same models on teacher-built and adjudicated labels. | Model rankings agree across the two label sets within their intervals, and no model trained on the teacher's labels exceeds the human agreement ceiling beyond its interval. |

Calibration that holds across schemas would turn the second regularity from an audit obligation into a property of the model. In the records available, in-domain recalibration left 0.024 to 0.89 of the raw calibration error, while transferred fits left as much as 0.96 of it [24, 25, 58, 116, 122]. Whether a temperature fitted on one schema holds on another is therefore an open problem. Joint uncertainty at low cost is a second opportunity. Proposition 5.1 shows that isolated answers identify a conjunction only within an interval, while the exact remedy, one Choice over the joint outcomes, grows exponentially with the number of questions. A dependence representation priced near the marginal calls would let a program combine answers without assuming independence.

Abstention that responds to missing evidence differs from abstention on low confidence [67], because a state can lack the fact a question needs while the model still reports a sharp distribution. The review band of Proposition 5.4 routes on probability alone. In Visual Jev, an open vision-language reimplementation, an answerability head detected missing evidence without improving decisions or selective prediction in that evaluation [166]. Contract versioning is open as well, since no study in the ledger tracks one deployment across versions of the hosted model, although a few evaluate jev-1.14.0 and one scores both versions on the same items [13]. Open checkpoints also move, since the main branch of Bespoke-Nimble-9B changed on 24 September 2026 to a checkpoint served at its unscaled default temperature [8].

Multimodal state extends the contract to images, and the evidence for it so far comes from open models alone. Visual Jev attaches a typed head to a vision-language backbone, and the full census lists further vision variants of open decision models (Section S5 of the electronic supplement). No study in the ledger runs an option-name swap in the sense of Proposition 5.5 or a cross-schema calibration audit on image states. Benchmark construction without teacher-consensus bias is the last problem. On the typed-decisions test split, a fresh teacher sample scored 0.735 against gold built from the other teacher samples, and the hosted model scored 0.727 [90], so scores near that ceiling are not separable. The Nimble holdout rests on synthetic labels [9], and neither reference set was adjudicated by people.

## 13 Conclusion

This survey has argued that decision models are a new contract rather than a new kind of intelligence, and has traced each element of that contract to one of five research streams. We defined the class and organized the open implementations by head, objective and interface contract. We then stated six propositions, each with its own premises, for any model honoring the contract, counted the ecosystem at a dated cutoff, and graded every included study for risk of bias. The ledger, with the option-name measurements of Section 8, supports three regularities. The contract travels, and robustness to option names belongs to each checkpoint, so it must be tested per model. Calibration belongs to the model and the task together, and is audited per schema. Escalation earns the saving when confidence ranks errors and the price gap is wide. A companion benchmark measures several decision models under one harness, and the pattern catalog, failure atlas, reporting standard and agenda turn these findings into guidance for software and research.

Several features of the evidence at the cutoff limit how far these conclusions extend. The literature spans the six weeks after the launch, and the census and the ledger describe the field at one cutoff. Because the training and the head of the hosted model are undisclosed, every statement about its architecture is inference. Within the ledger, the first regularity rests on 9 parity-controlled comparisons for its accuracy gap, all of zero-shot models, and the lead of task-trained models comes only from studies whose authors built those models. We added the split by comparator class after the first synthesis of this version, although a blind recoding agreed closely with it. The third regularity rests on 23 cascade studies with different cost definitions. The companion benchmark adds a swap test of every decision model and cascades under one cost definition, but its typed-decisions labels

come from a teacher model, so it addresses only part of this gap. Only 3 studies reached a low overall risk of bias, and the grades depend strongly on how the exposure of public benchmarks is rated (Section S3). Part of the evidence rests on reference labels from teacher models, synthetic procedures or a vendor-chosen consensus of language models rather than human adjudication. We also formulated the three regularities after extraction, as summaries of the graded rows. The authors' own studies are graded under the rules applied to every other ledger row, and removing them leaves the zero-shot gap at −0.7 points and the frontier gap at −3.3 points. Versions also move under unchanged names, and the companion benchmark measures a newer Bespoke-Nimble-9B revision than the ledger rows describe.

Several of these bounds point to further studies, some made concrete in the agenda of Section 12. Repeated sweeps against later cutoffs, released as new minor versions of the companion repository, can test whether the regularities hold as the literature grows. Independent replication of the authors' own measurements, which the versioned census, ledger and harness are built to support, would remove the remaining dependence on their judgment. Further parity-controlled comparisons under a shared cost definition would narrow the ranges behind the first and third regularities, and the contract-versioning study of Table 7 would track hosted and open checkpoints across releases. Benchmarks labeled by human adjudication on items that postdate the models would address the exposure ratings and the teacher-consensus bias. Until the schema-transfer study settles whether calibration can become a property of the model, each deployment should audit calibration on every schema it uses.

## References


[1] 2006. *An Introduction to Copulas.* Springer New York. doi:10.1007/0-387-28678-0

[2] Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning* 16, 4 (2023), 494–591. doi:10.1561/2200000101

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

[5] Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic Calibration of Long-Form Generations. arXiv:2404.00474 [cs.LG] https://arxiv.org/abs/2404.00474

[6] 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

[7] M. Bergamaschi Ganapini, M. Campbell, F. Fabiano, L. Horesh, J. Lenchner, A. Loreggia, N. Mattei, F. Rossi, B. Srivastava, and K. B. Venable. 2025. Fast, slow, and metacognitive thinking in AI. *npj Artificial Intelligence* 1, 1 (2025). doi:10.1038/s44387-025-00027-5

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

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

[10] Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jon Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. 2020. Thinking Fast and Slow in AI. arXiv:2010.06002 [cs.AI] https://arxiv.org/abs/2010.06002

[11] 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

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[13] 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

[14] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176 [cs.LG] https://arxiv.org/abs/2305.05176

[15] 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

[16] 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

[17] 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

[18] C. K. Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* EC-6, 4 (1957), 247–254. doi:10.1109/tec.1957.5222035

[19] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] https://arxiv.org/abs/1706.03741

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

[21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168

[22] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385 (2024), e078378. doi:10.1136/bmj-2023-078378

[23] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G.M. Moons. 2015. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *European Urology* 67, 6 (2015), 1142–1151. doi:10.1016/j.eururo.2014.11.025

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

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

[26] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Learning with Rejection. In *Lecture Notes in Computer Science*. Springer International Publishing. doi:10.1007/978-3-319-46379-7_5

[27] Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. arXiv:2507.16806 [cs.LG] https://arxiv.org/abs/2507.16806

[28] 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

[29] 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

[30] Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 295–302. doi:10.18653/v1/2020.emnlp-main.21

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186. doi:10.18653/v1/n19-1423

[32] David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton. 2022. Language Model Cascades. arXiv:2207.10342 [cs.CL] https://arxiv.org/abs/2207.10342

[33] Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2025. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models. In *Proceedings of Machine Learning and Systems (MLSys)*, Vol. 7. https://arxiv.org/abs/2411.15100

[34] 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

[35] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv:2404.04475 [cs.LG] https://arxiv.org/abs/2404.04475

[36] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research* 11 (2010), 1605–1641.

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

[38] 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

[39] Jonathan St. B. T. Evans. 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology* 59, 1 (2008), 255–278. doi:10.1146/annurev.psych.59.103006.093629

[40] Francesco Fabiano, Marianna B. Ganapini, Andrea Loreggia, Nicholas Mattei, Keerthiram Murugesan, Vishal Pallagani, Francesca Rossi, Biplav Srivastava, and K. Brent Venable. 2025. Thinking Fast and Slow in Human and Machine Intelligence. *Commun. ACM* 68, 8 (2025), 72–79. doi:10.1145/3715709

[41] Marianna Bergamaschi Ganapini, Murray Campbell, Francesco Fabiano, Lior Horesh, Jon Lenchner, Andrea Loreggia, Nicholas Mattei, Francesca Rossi, Biplav Srivastava, and Kristen Brent Venable. 2021. Thinking Fast and Slow in AI: the Role of Metacognition. arXiv:2110.01834 [cs.AI] https://arxiv.org/abs/2110.01834

[42] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[43] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 4878–4887. https://arxiv.org/abs/1705.08500

[44] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6577–6595. doi:10.18653/v1/2024.naacl-long.366

[45] 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

[46] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. arXiv:1706.04599 [cs.LG] https://arxiv.org/abs/1706.04599

[47] 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

[48] Hacker News. 2026. *Introducing System One Models and Jev (launch discussion thread).* Retrieved 2026-09-24 from https://news.ycombinator.com/item?id=49717558

[49] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838. doi:10.1038/s43588-023-00527-x

[50] 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

[51] 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

[52] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: a survey. *Machine Learning* 113, 5 (2024), 3073–3110. doi:10.1007/s10994-024-06534-x

[53] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1610.02136

[54] D. G. Horvitz and D. J. Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *J. Amer. Statist. Assoc.* 47, 260 (1952), 663–685. doi:10.1080/01621459.1952.10483446

[55] 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

[56] 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

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

[58] 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

[59] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674 [cs.CL] https://arxiv.org/abs/2312.06674

[60] 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

[61] 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

[62] Wittawat Jitkrittum, Neha Gupta, Aditya Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. 2023. When Does Confidence-Based Cascade Deferral Suffice?. In *Advances in Neural Information Processing Systems 36*. 9891–9906. arXiv:2307.02764 doi:10.52202/075280-0431

[63] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs.CL] https://arxiv.org/abs/2207.05221

[64] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[65] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. arXiv:2509.04664 [cs.CL] https://arxiv.org/abs/2509.04664

[66] Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 160–171. arXiv:2311.14648 doi:10.1145/3618260.3649777

[67] Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective Question Answering under Domain Shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5684–5696. doi:10.18653/v1/2020.acl-main.503

[68] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4, 9 (2023), 100804. doi:10.1016/j.patter.2023.100804

[69] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. Active Testing: Sample-Efficient Model Evaluation. arXiv:2103.05331 [stat.ML] https://arxiv.org/abs/2103.05331

[70] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. arXiv:1910.12656 [cs.LG] https://arxiv.org/abs/1910.12656

[71] Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified Uncertainty Calibration. arXiv:1909.10155 [cs.LG] https://arxiv.org/abs/1909.10155

[72] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal Prediction with Large Language Models for Multi-Choice Question Answering. arXiv:2305.18404 [cs.CL] https://arxiv.org/abs/2305.18404

[73] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lj Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 1755–1797. arXiv:2403.13787 doi:10.18653/v1/2025.findings-naacl.96

[74] 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

[75] Latent Space (swyx). 2026. *System One models for Prod not God*. Retrieved 2026-09-24 from https://www.latent.space/p/jev

[76] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880. doi:10.18653/v1/2020.acl-main.703

[77] 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

[78] 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

[79] 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

[80] Hengyu Li. 2026. KITE: Scaling Jev Population Experiments with Sparse Flagship Calibration. arXiv:2609.27535 [cs.MA] https://arxiv.org/abs/2609.27535

[81] 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

[82] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal

Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://arxiv.org/abs/2211.09110

[83] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. arXiv:2305.20050 [cs.LG] https://arxiv.org/abs/2305.20050

[84] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2023. SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks. In *Advances in Neural Information Processing Systems 36*. 23813–23825. arXiv:2305.17390 doi:10.52202/075280-1034

[85] 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

[86] 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

[87] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. 6107–6117. doi:10.1145/3711896.3736569

[88] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* (2020), m3164. doi:10.1136/bmj.m3164

[89] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2511–2522. arXiv:2303.16634 doi:10.18653/v1/2023.emnlp-main.153

[90] LocalLLaMA. 2026. *LocalLLaMA/typed-decisions*. Retrieved 2026-09-24 from https://huggingface.co/datasets/LocalLLaMA/typed-decisions

[91] Luni. 2026. *Luni/laya-jev-benchmark*. Retrieved 2026-09-24 from https://huggingface.co/datasets/Luni/laya-jev-benchmark

[92] Nandakishor M. 2025. Confidence-Aware Routing for Large Language Model Reliability Enhancement: A Multi-Signal Approach to Pre-Generation Hallucination Mitigation. arXiv:2510.01237 [cs.CL] https://arxiv.org/abs/2510.01237

[93] Nandakishor M. 2025. SalesRLAgent: A Reinforcement Learning Approach for Real-Time Sales Conversion Prediction and Optimization. arXiv:2503.23303 [cs.LG] https://arxiv.org/abs/2503.23303

[94] 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

[95] MachineFi. 2026. *Trio-Spark Model Card*. https://huggingface.co/MachineFi/Trio-Spark

[96] David Madras, Toniann Pitassi, and Richard Zemel. 2017. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. arXiv:1711.06664 [stat.ML] https://arxiv.org/abs/1711.06664

[97] Mapika. 2026. *Mapika/decider-2b*. Retrieved 2026-09-24 from https://huggingface.co/Mapika/decider-2b

[98] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A Holistic Approach to Undesired Content Detection in the Real World. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (2023), 15009–15018. doi:10.1609/aaai.v37i12.26752

[99] 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

[100] Microsoft. 2026. *Introducing Microsoft-Decision-1*. https://commandline.microsoft.com/microsoft-decision-1-model-foundry/

[101] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the Calibration of Modern Neural Networks. arXiv:2106.07998 [cs.LG] https://arxiv.org/abs/2106.07998

[102] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229. doi:10.1145/3287560.3287596

[103] Karel G.M. Moons, Robert F. Wolff, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. 2019. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine* 170, 1 (2019), W1–W33. doi:10.7326/m18-1377

[104] Yasmin Moslem and John D. Kelleher. 2026. Dynamic Model Routing and Cascading for Efficient LLM Inference: A Survey. arXiv:2603.04445 [cs.NI] https://arxiv.org/abs/2603.04445

[105] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. arXiv:2006.01862 [cs.LG] https://arxiv.org/abs/2006.01862

[106] Allan H. Murphy. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* 12, 4 (1973), 595–600. doi:10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

[107] Andrew Y. Ng and Michael I. Jordan. 2001. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In *Advances in Neural Information Processing Systems 14*. https://proceedings.neurips.cc/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html

[108] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. RouteLLM: Learning to Route LLMs with Preference Data. arXiv:2406.18665 [cs.LG] https://arxiv.org/abs/2406.18665

[109] OpenRouter. 2026. *Jev Documentation: TypeSafe Decision Model on OpenRouter*. Retrieved 2026-09-24 from https://openrouter.ai/docs/guides/community/jev

[110] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems 35*. 27730–27744. arXiv:2203.02155 doi:10.52202/068431-2011

[111] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*. https://arxiv.org/abs/1906.02530

[112] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* (2021), n71. doi:10.1136/bmj.n71

[113] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (2015). doi:10.1609/aaai.v29i1.9602

[114] 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

[115] Palmer, Jared. 2026. *jaredpalmer/kev-0.8b*. Retrieved 2026-09-24 from https://huggingface.co/jaredpalmer/kev-0.8b

[116] Palmer, Jared. 2026. *jaredpalmer/kev-9b*. Retrieved 2026-09-24 from https://huggingface.co/jaredpalmer/kev-9b

[117] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems 37*. 68772–68802. arXiv:2404.13076 doi:10.52202/079017-2197

[118] Perplexity. 2026. *pplx-decider-v1.1-27b Model Card*. https://huggingface.co/perplexity-ai/pplx-decider-v1.1-27b

[119] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). arXiv:2003.12206 [cs.LG] https://arxiv.org/abs/2003.12206

[120] John C. Platt. 2000. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (Eds.). MIT Press, 61–74.

[121] 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

[122] 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

[123] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44. doi:10.1145/3351095.3372873

[124] Rauch. 2026. *post on X about Jev latency in Vercel AI Gateway*. Retrieved 2026-09-24 from https://x.com/rauchg/status/2100307962262872105

[125] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3980–3990. doi:10.18653/v1/d19-1410

[126] 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

[127] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, and Melanie J Calvert. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* (2020), m3210. doi:10.1136/bmj.m3210

[128] Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging Large Language Models for Multiple Choice Question Answering. arXiv:2210.12353 [cs.CL] https://arxiv.org/abs/2210.12353

[129] 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

[130] Haggai Roitman. 2026. The Hitchhiker's Guide to System 1 Models: Fast Decision-Making Beyond LLMs. arXiv:2610.33775 [cs.AI] https://arxiv.org/abs/2610.33775

[131] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems 36*. 68539–68551. arXiv:2302.04761 doi:10.52202/075280-2997

[132] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 255–269. doi:10.18653/v1/2021.eacl-main.20

[133] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2026. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. *Comput. Surveys* 58, 3 (2026), 1–38. doi:10.1145/3744238

[134] 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

[135] Keith E. Stanovich and Richard F. West. 2002. Individual Differences in Reasoning: Implications for the Rationality Debate? In *Heuristics and Biases*. Cambridge University Press. doi:10.1017/cbo9780511808098.026

[136] Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, Oleksandr Lukashov, Alexander Yavorskyi, and Mykyta Yaroshenko. 2025. GLiClass: Generalist Lightweight Model for Sequence Classification Tasks. arXiv:2508.07662 [cs.LG] https://arxiv.org/abs/2508.07662

[137] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. arXiv:2009.01325 [cs.CL] https://arxiv.org/abs/2009.01325

[138] Joshua Strong, Emma Sun, Harry Rogers, Helen Higham, and Alison Noble. 2025. Learning to Defer: A Survey. doi:10.5281/zenodo.17843044

[139] 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

[140] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1218–1236. arXiv:2408.02442 doi:10.18653/v1/2024.emnlp-industry.91

[141] Lijuan Tang and Yuemeng Zheng. 2026. Typed Decision Models: An Early Evidence Audit and Evaluation Checklist. arXiv:2609.32160 [cs.CL] https://arxiv.org/abs/2609.32160

[142] 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

[143] TypeSafe AI. 2026. *AI Primer*. Retrieved 2026-09-24 from https://docs.typesafe.ai/introduction/machine-learning-primer

[144] TypeSafe AI. 2026. *API Reference*. Retrieved 2026-09-24 from https://docs.typesafe.ai/api

[145] TypeSafe AI. 2026. *Box*. https://docs.typesafe.ai/primitives/box.md

[146] TypeSafe AI. 2026. *Confidence*. Retrieved 2026-09-24 from https://docs.typesafe.ai/confidence

[147] TypeSafe AI. 2026. *Introducing System One Models & Jev*. Retrieved 2026-09-24 from https://typesafe.ai/blog/introducing-system-one-models-and-jev

[148] TypeSafe AI. 2026. *Introduction*. Retrieved 2026-09-24 from https://docs.typesafe.ai/introduction

[149] TypeSafe AI. 2026. *Jev 1.13 jaggedness*. Retrieved 2026-09-24 from https://docs.typesafe.ai/model-jaggedness/jev-1.13

[150] TypeSafe AI. 2026. *Models*. https://docs.typesafe.ai/models.md

[151] TypeSafe AI. 2026. *Patterns*. Retrieved 2026-09-24 from https://docs.typesafe.ai/patterns

[152] TypeSafe AI. 2026. *Score*. Retrieved 2026-09-24 from https://docs.typesafe.ai/primitives/score

[153] TypeSafe AI. 2026. *System One.* Retrieved 2026-09-24 from https://docs.typesafe.ai/concepts/system-one

[154] TypeSafe AI. 2026. *Workflow evals.* Retrieved 2026-09-24 from https://evals.typesafe.ai/

[155] Clovis Varangot-Reille, Christophe Bouvard, Mathieu Ciancone, Antoine Gourru, Marion Schaeffer, and François Jacquenet. 2026. Doing More with Less: A Survey on Routing Strategies for Resource Optimisation in Large Language Model-Based Systems. *Journal of Artificial Intelligence Research* 86 (2026). doi:10.1613/jair.1.19801

[156] Paul Viola and Michael Jones. 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. I–511–I–518. doi:10.1109/CVPR.2001.990517

[157] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2526–2547. arXiv:2412.13663 doi:10.18653/v1/2025.acl-long.127

[158] 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

[159] Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. arXiv:2307.09702 [cs.CL] https://arxiv.org/abs/2307.09702

[160] Willison, Simon. 2026. *Jev introduces a new shape of LLM.* Retrieved 2026-09-24 from https://simonwillison.net/2026/Sep/21/jev/

[161] 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

[162] 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

[163] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=gjeQKFxFpZ

[164] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736 [cs.CL] https://arxiv.org/abs/2410.02736

[165] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3912–3921. doi:10.18653/v1/d19-1404

[166] 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

[167] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 694–699. doi:10.1145/775047.775151

[168] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 5364–5376. arXiv:2311.08526 doi:10.18653/v1/2024.naacl-long.300

[169] Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. ShieldGemma: Generative AI Content Moderation Based on Gemma. arXiv:2407.21772 [cs.CL] https://arxiv.org/abs/2407.21772

[170] 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

[171] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690 [cs.CL] https://arxiv.org/abs/2102.09690

[172] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=shr9PXz7T0

[173] 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

[174] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291. doi:10.1162/coli_a_00502