RIGHT NUMBER, WRONG QUESTION: LLM PROBABILITY SUBSTITUTION IN PARTIAL SEARCHES

Anonymous authors
Paper under double-blind review

ABSTRACT

Suppose an agent fails to find the target in a partial search and must report the chance the target still exists in the unsearched share. How do language models respond to a required probability field when the prompt does not contain the base rate for the search target? In this setting, when the prompt states the search's coverage as the share of records each searched location holds, we find that 22 of 46 substitute a different probability event for the requested posterior. Supplying the base rate removes the substitution in most models, and asking the model for its assumed prior removes the substitution in half of the models scored. Even when the prompt asks for the prior and for the event the number describes, relabeling the probability field as "the number you report" brings the substitution back. The substitution is a default rather than a misunderstanding. When asked in separate inference calls for the probability that the target exists and for the probability that the search missed it, 18 of 30 models answer the same number to both questions. This two-call test detects 19 of 20 substituting models and none of the seven that report the posterior, so a pipeline can screen a model before trusting its probabilities. The substitution persists even when some models run the search themselves with tools. A calibration score cannot tell which event this default probability refers to, and no recalibration of the reported probability can correct to the posterior at every base rate. We simulate a thresholding pipeline on two datasets and show how thresholding on these LLM-reported probabilities leads to silent failures in downstream applications at low base rates. As self-reported model probabilities are commonly used to make decisions and orchestrate agent pipelines, careful evaluations should test what events those probabilities are actually estimating.

1 INTRODUCTION

How do large language models (LLMs) report probabilities when the prompt is underspecified? Does the model decline, assume the missing inputs, or return a poorly calibrated estimate? In the case of estimating the probability that a search target exists, we find that several models instead substitute a different event for the requested one. Models report the probability that their search would have missed the target, which is a correct probability for the wrong probability event. Asked which event the number describes, most models correctly identify that event. Because the number is a true probability of that other event, a calibration score alone does not flag it, and a pipeline that thresholds it applies the threshold to the wrong question.

We consider the setting of partial searches where the agent fails to find the target and only searches part of the available search space. After the failed search, the agent reports the probability that the target still exists in the unsearched remainder. For example, the agent fails to find a user's parking permit after only searching 32% of a company's records and needs to report the probability the target exists in the remainder of the records. This probability of target existence depends on two inputs: the coverage (known and given in the prompt) and the base rate of the target a priori (not given). If the LLM reports the miss likelihood, the unsearched share of the space (0.68 in the example), for the requested posterior, we call this a *substitution*. This substituted event presupposes guaranteed target existence somewhere in the search space and varies with coverage. We find that 22 of 46 LLMs commit this substitution of probability events in the partial search setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We systematically investigate the extent of this substitution behavior and the conditions under which it appears. We vary the coverage δ across episodes while keeping the search evidence fixed, and we ask different questions and require different output fields about the same failed search. In most experiments the prompt contains a search log. In two settings the model runs the search itself with tool calls, and the substitution persists in some models. The substitution is not an inability or a misunderstanding of the question. Models are able to name the event they report, and they give the posterior and the miss likelihood the same number less often when one reply asks for both, but they still commit the substitution by default. How often the substitution occurs depends on the output field schema. The substitution is specific to inferring a probability number (Fig. 1). Allowed to decline instead, about a third of models decline most partial searches, but few models decline enough to avoid all suppressions. Supplying the base rate corrects the report in most models. Asking the model for the prior it assumes also corrects the report, and the substitution returns when the probability field is reworded while the prior is asked for. Downstream, a pipeline that thresholds the substituted number skips the record without review, and no calibration check on the number can tell that it is a probability of a different event (Sec. 3).

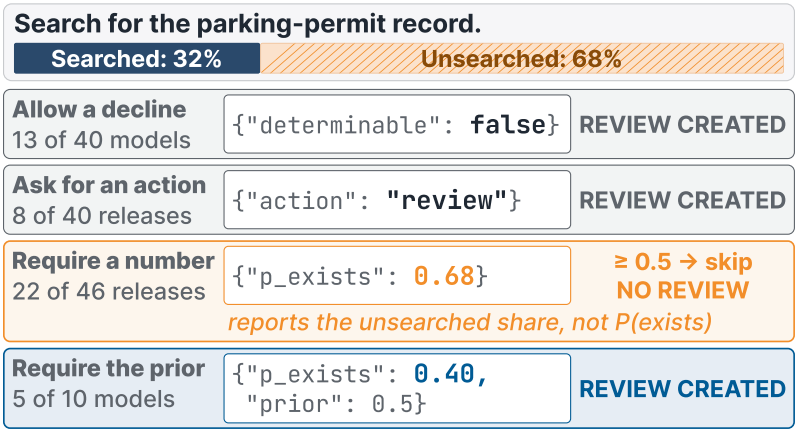


Figure 1: **How output field schema changes downstream decisions in partial search.** Suppose an agent searches 32% of records and fails to find the target. When allowed to decline, 13 of 40 models send most of their partial searches up for manual review, and the median model declines 30% of them. However, when required to output a probability number, 22 of the 46 substitute the event $P(F \mid E)$ for $P(E \mid F)$ and a 0.5 threshold skips the record without review. When required to also give the prior, five of the ten models scored report the posterior at that prior, here 0.40. Eight of the 40 releases decline often enough, in 88% or more of cases, that no record is suppressed (App. H).

Our contributions are as follows: 1) We find that probability substitution is a default in about half of the LLMs we scored (22 of 46 releases from ten companies, the majority rule in 6 of 17 families) under an underspecified prompt, governed by the output field schema rather than by the models' ability, and that the substitution persists in some models when they run the search themselves and on real records (Sec. 3). 2) We construct a benchmark whose search coverage is known by design, so that varying the coverage identifies which event a reported probability tracks and separates event identification from calibration, and we apply the benchmark to 46 LLMs from ten companies (Sec. 2; App. A). 3) We show the consequence and the remedy. A calibration score cannot tell which event the number describes, and no function of the number alone gives the posterior at every base rate, so thresholding pipelines fail silently; a two-call check detects the substituting models, and computing the posterior outside the model removes the substitution's cost (Sec. 3; Apps. G and H).

2 Identifying Probability Response Rules

The standard search model requires two inputs: the detection probability δ, which we call the search coverage throughout, and the base rate π (Stephens, 2011; Sober, 2009). Let E be the event that the target exists and F the event that the search misses it. The target, if it exists, is in one of N locations, the set C (six in every episode of the benchmark). Location $\ell \in C$ holds a *location share* $w_\ell > 0$ of such records, with $\sum_{\ell \in C} w_\ell = 1$, and a search that reads the set S in full has coverage

$\delta(S) = \sum_{\ell \in S \cap C} w_\ell$. A nonexistent target is never found, so $P(F \mid \neg E) = 1$, and the search misses an existing target with probability $P(F \mid E) = 1 - \delta$. Therefore, the posterior that the target exists is

$$p^*(\delta, \pi) \;=\; P(E \mid F) \;=\; \frac{\pi(1-\delta)}{(1-\pi)+\pi(1-\delta)} \;=\; \frac{\pi(1-\delta)}{1-\pi\delta}, \tag{1}$$

where $1-\delta$ is the miss likelihood and π the base rate, how common such targets are before the search. A *prior* is a value of π that a model assumes or is given. We write $p^*(\delta, \pi)$ when the posterior is treated as a function of coverage and base rate, and $P(E \mid F)$ when we mean the probability of the event itself. A model that reports the miss likelihood in place of the posterior makes the *substitution*. A model's whole output to one episode is its *response*. A response can have several *fields*, one for each value it must give, and the probability a model writes in the probability field is its *report*. A *response rule* is a fixed mapping from the search evidence in the prompt (the search log and the location shares, which the coverage δ summarizes) to a reported probability. Below, *the rule* means the $1 - \delta$ response rule.

The miss likelihood depends only on the search, whereas the posterior depends on the search and on the base rate. A prompt that gives the shares and the searched set determines δ and leaves π unspecified, so the posterior is underdetermined. The propositions in App. G evaluate a report against every base rate in $[0.1, 0.9]$ rather than one. In the results we compare responses with each rule's curve, and take $\pi = 0.5$ for the posterior rule.

We test three reference rules, whose predicted curves are:

- $1 - \delta$ *rule*: outputs $1 - \delta$ with slope -1 in δ;
- *0.5-prior posterior*: outputs $p^*(\delta, \frac{1}{2}) = (1-\delta)/(2-\delta)$ with slope $-1/(2-\delta)^2$;
- *Counting rule*: builds the coverage estimate k/N from the k of the N candidate locations it treats as searched, then outputs $1 - k/N$. A model that counts the same way but reports the posterior at k/N follows the *counting posterior*.

Under a supplied prior, only the posterior is correct. The $1 - \delta$ rule reports the miss likelihood, $P(F \mid E) = 1 - \delta$, which is not the probability that the target exists. The two coincide only when $\pi = 1/(1+\delta)$ (Good, 1950), so no single fixed prior fits the report at every coverage. The report is a well-formed probability that decreases as the coverage increases, so the number alone does not show that the base rate was never used. We use $\pi = 0.5$ as the reference posterior because episodes do not give a base rate and 0.5 is the neutral choice models most often adopt (App. D.6). We identify the substitution by asking about both events of one search and by the event field; this prior only fixes where the curves are drawn.

As in Fig. 1, an agent checks whether an employee has a parking-permit record. The searched set is known to hold 32% of such records (given in the prompt), so $\delta = 0.32$. At $\pi = 0.5$ the posterior is $p^*(0.32, \frac{1}{2}) = (0.5 \times 0.68)/(0.5 + 0.5 \times 0.68) = 0.405$, so $P(E \mid F) = 0.405$, and a model on the $1 - \delta$ rule reports $P(F \mid E) = 1 - \delta = 0.68$ instead. Base-rate neglect in human judgment occurs when the prior is given (Bar-Hillel, 1980), as does the prosecutor's fallacy (Thompson & Schumann, 1987). In our setting no base rate is given, so the model returns a number determined by the search evidence alone.

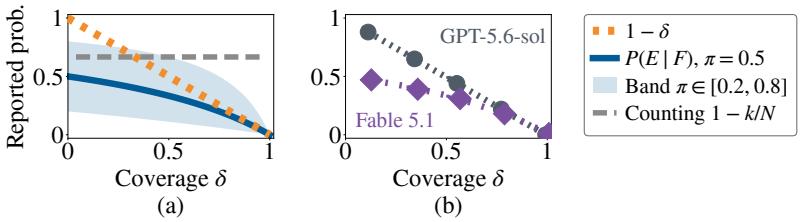


Figure 2: **Varying δ discriminates between candidate response rules.** **(a)** The three rules as curves over the coverage δ: the miss likelihood $1 - \delta$, the posterior at $\pi = 0.5$ with its band over $\pi \in [0.2, 0.8]$, and the counting rule $1 - k/N$ at $k = 2$ of $N = 6$, drawn only where the count is defined. **(b)** Per-level medians of two models: GPT-5.6-sol follows $1 - \delta$ and Fable 5.1 the 0.5-prior posterior. All model curves are plotted in Fig. 9.

The three rules produce different curves over δ (Fig. 2). Two rules may agree at one δ, but their slopes are different, so the response curve identifies the rule (Lemma 1, App. G). We place the δ points where the candidates differ most (Myung & Pitt, 2009). We ask which search event each response refers to, not how well calibrated the number is (Tian et al., 2023; Xiong et al., 2024).

The benchmark and the criterion. If the model had to estimate the coverage, we could not tell a model reporting the miss likelihood from one reporting a posterior at a different coverage. Therefore, every prompt gives the coverage. Each episode gives the model a failed search and asks for the posterior $P(E \mid F)$ as a JSON field. In most episodes the prompt contains the search log. In two settings the model runs the search, choosing the sections it opens or issuing grep queries, and δ comes from what it searched (Sec. 3.1). A bug form asks whether a bug is present after tests over part of a program's modules (Sec. 3.1). The locations named in the prompt set the coverage δ, and the generator plants the records, so the search has failed and its coverage is known by construction (App. A.1). Episodes do not give a base rate unless an experiment adds one. Two external datasets, clinical records and enterprise documents, supply the episodes for the pipeline of Sec. 3.3, where present and absent records are mixed at a chosen base rate. We work mainly at four partial-coverage levels, with the empty and exhaustive searches as edge cases and external data in Sec. 3.3. We exclude four models that did not return a parsable number on any query, one of which refused the prompt (App. I).

Additional fields. We add three fields beside the probability field; each field rules out one explanation of the number. A *second question* asks for the miss likelihood $P(F \mid E)$ in its own prompt, so a model that means the miss likelihood can say so. An *event label* field asks which event the probability describes, separating a model that substitutes from one that has simply named the wrong event. A *prior* field supplies the base rate or requires it, separating a missing input from an inability to use one (App. A, with each form's fields in Table 17). Each *output schema* is a one-turn JSON reply with the probability field and some of these additions (Table 17 lists the schemas used). The event label comes in two forms: a field in the same reply (one turn), or a second turn about the event the earlier probability describes (two turns).

Assigning a rule. Varying δ across its range with the search environment kept fixed gives each candidate rule a different curve, so the responses identify one rule (Fig. 2). We assign a model a rule when at least 60% of its polled partial-coverage responses fall within 0.02 of that rule's curve, over 122 to 1,478 responses per model. We call a model *mixed* when no rule meets this criterion. Of the 17 mixed models, 13 are closer to a posterior at one fitted prior than to $1-\delta$ (Table 5). We score the event and prior fields against the same tolerance. A response within the tolerance of two rules is not counted for either (App. A.4, Table 7).

3 Results

3.1 LLMs exhibit probability substitution behavior in partial search. Asked for a posterior probability $P(E \mid F)$ when the base rate is withheld, 22 of the 46 LLMs fill the field with the miss likelihood $1-\delta$. When asked to derive the probability in the reply first, models such as GPT-5.6-sol identify the substituted event; in an episode at $\delta = 0.36$:

> "Strictly, an absolute existence probability would require a prior; using the remaining filing-share coverage as the requested estimate gives $1 - 0.36 = 0.64$."

We also explicitly ask the model which event its probability number refers to, in one turn (a required field) and in two turns (a second turn after the reply). In the one-turn schema, whose probability field reads "the number you report", 256 of the 265 $1-\delta$ responses from the ten scored models name the unsearched share or the miss likelihood in the event field, and only 9 claim the posterior (App. C.5). Therefore, most models are able to tell which quantity their self-reported probability refers to and are not misinterpreting the prompt or question asked. To test whether $1-\delta$ is a default that the model applies unless the prompt requires otherwise, we change the output fields while keeping the search fixed. In Fig. 3, we test other single-turn output schemas (Sec. 2), where the event and prior are required fields in addition to the probability number. We find that the substitution is more common when the probability field is redefined as "the number you report" (Table 15), so the substitution is sensitive to the wording of the probability field rather than to which other fields the reply asks for. In a two-turn prompt where the event is asked after the probability, only 7 out of 10 models name their $1-\delta$ event correctly (App. C.5). The two-turn design fixes the number before the event is

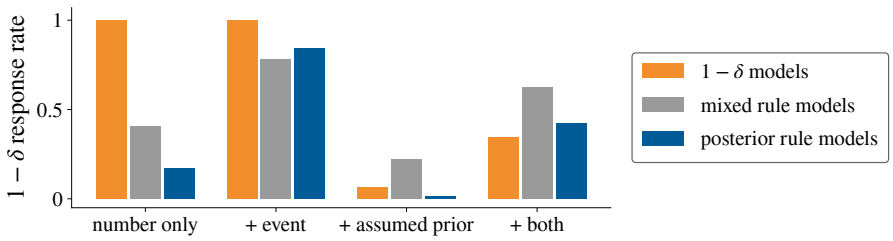


Figure 3: **Effect of output schema on substitution behavior.** Each bar is the median, over all models following one of the three rules, of the percent of replies matching the miss likelihood $1-\delta$. The first group (number only) on the left only asks the model for the posterior probability number field. The second group (number + event) also asks the model which event its number describes by rewording the number field as "the number you report". The third group asks for the probability number and what prior the model assumes while the last group on the right asks for all three outputs. We see that adding the event field with the number rewording causes models (even those following the mixed or the posterior rule) to increase in $1-\delta$ substitution behavior, while asking for a prior mitigates the behavior.

asked, and tests naming without letting the field change the number. We do not know why the other three name existence. The redefined probability field that comes with the one-turn event field, not the event field itself, causes some models, such as Gemini 3.1 Pro, to substitute $1-\delta$.

Therefore, the substitution can be induced in models that do not initially substitute $1-\delta$ by rewording the required number field. Furthermore, asking the model to supply the missing base rate input itself restores half of the ten scored models to the normal posterior behavior at the assumed prior (Fig. 8), while explicitly giving the prior in the prompt restores 12 of 18 models to the posterior (App. D.1). An event field beside a probability field that still asks about existence barely changes the posterior responses; rewording that field as "the number you report" brings the substitution back, and the number of responses at the posterior drops from 418 to 266 of 448 (totals and per model in App. D.3).

We find that default substitution behavior is specific to a probability output field over searches where coverage can be computed from the prompt. If asked instead for an expected count or value after a partial audit that did not find any of the items counted, the 11 models tested answer a count or value of 0 in 2,601 of 2,611 responses. If the prompt does not give any search coverage (only retrieved articles that never mention the target), most models answer below the base rate itself (the base rate measured on that dataset, which the prompt does not state). Given the coverage as a running percentage (a single coverage figure updated after each step) in place of the locations' shares, models do not consistently report the miss likelihood. We do not know the cause of this difference between the share table and the running percentage (App. B). Substitution behavior is also confined to the number output field. For a decline output, about a third of the models decline most partial searches (Table 13); a decline is a structured output in place of the number that sends the search to manual review.

Substitution behavior also transfers, in some models, to agentic searches with tool calls. GPT-5.5, choosing which sections of a 511,962-document enterprise corpus to search, reports the miss likelihood of the stated shares in 63 of 64 episodes (App. B). In a live search for a named function in a generated code repository with a grep tool, GPT-5.5 and Grok 4.20 Reasoning, two of the five models on the $1-\delta$ rule we could score, report the miss likelihood of the search they ran at every budget. However, Haiku 4.5, Nova Pro and Sonnet 5 report less (Table 9). When asked whether a bug is present after tests over part of a program, all six models on the $1-\delta$ rule report the untested share of the program as that probability in 95 to 100% of their replies (Table 10). These settings include real document collections and live tool use, so the substitution is not specific to the generated benchmark. On MIMIC's hospital medication tables, real clinical records with a measured base rate of 0.72, GPT-5.5's most common response after a miss is $1-\delta$ for that table (Table 8).

3.2 Causes and interventions for substitution behavior. Default substitution behavior is not an inability to comprehend the prompt or calculate the coverage. Models are able to correctly count

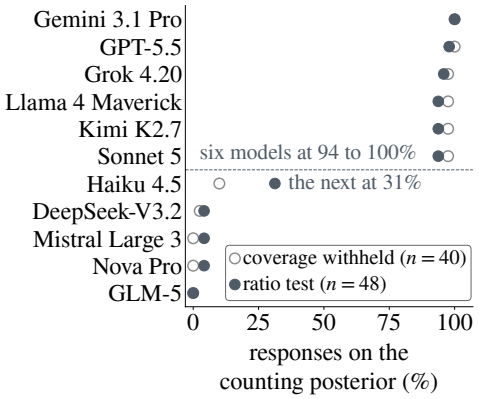


Figure 4: **Given a list of locations but no coverage number, six of the 11 models count the searched locations and report the posterior at that count.** Each point is one model's share of responses at the posterior computed from the count of searched locations, filled when the list keeps the same searched-to-total ratio and hollow when the coverage figure is withheld. The six score 94 to 100% and the next model 31%; doubling the list at the same coverage does not change the share.

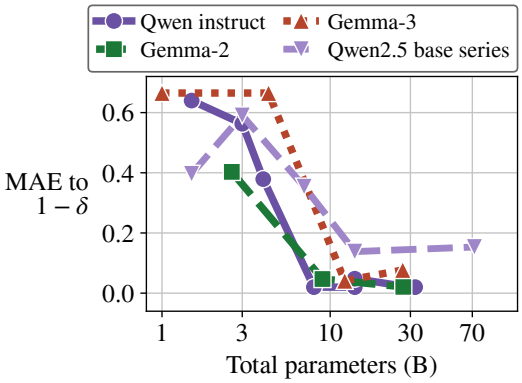


Figure 5: **The reported number moves toward $1-\delta$ as model scale grows.** Distance to $1-\delta$ on the coverage-stated condition against total parameters, within three open-weight instruct families; the Qwen2.5 base series (dashed) decreases later and incompletely. Coverage estimation by counting appears at larger scale (Fig. 13).

the number of locations searched and compute coverage from provided distributions (six of the 11 models, given the base rate and a list of places without a coverage figure, report the counting posterior of Sec. 2; Fig. 4).

A linear probe (Alain & Bengio, 2016) recovers the stated base rate from the activations of Qwen3-32B and Gemma-2-9B-it. Splicing the activations of a prompt that states the base rate into one that withholds it, an activation patch (Meng et al., 2022), changes Qwen3-32B's response to the posterior in 79 of 128 responses at one block, and the result replicates on a second seed (App. F). The patch works from the input layer on and does not localize the conversion; other models are untested.

We also try fine-tuning three models (Gemma-2-9B-it, OLMo-2-13B-Instruct and Qwen2.5-14B-Instruct) with a LoRA reward for posterior responses on templates that state the prior, testing on held-out templates (App. F.5). Fine-tuning produces posterior responses only when the prompt carries the training prompts' coverage line. No model reports the posterior in 60% of replies on a prompt it never saw (Gemma-2-9B-it, 14.6%, against 83 to 90% with that line added). We also test whether the behavior is a result of scaling training data or model size. In Fig. 5, across three open-weight instruct families the distance of the reported number from $1-\delta$ shrinks with parameter count, and a base-model series shrinks the distance later and less completely. Therefore, the default appears at smaller scales than coverage estimation itself, and is sharper after post-training. Among frontier closed models the rule is not universal, and five of the seven posterior-rule models are Anthropic's (Table 7).

The default can be detected from the replies alone. We ask for the probability that the target exists, $P(E \mid F)$, and for the probability that the search would have missed it if it exists, $P(F \mid E) = 1-\delta$. When each question is asked in its own call with no shared history, 18 of the 30 models passing a calculation check give both events the same number, so the number does not depend on the event named (Fig. 6; App. C.1). When both questions are asked in one reply, most models on the $1-\delta$ rule give the two events the same number less often (Fig. 12), but a reply that departs from $1-\delta$ does not necessarily give the posterior. Only about a third of the departing replies give the posterior at a prior of 0.5 (App. C.4). The posterior-rule models assume a prior of 0.5 on the generated benchmark. The external datasets (App. B) test the transfer of the substitution, and the pipeline simulation of Sec. 3.3 supplies the base rate instead of asking the model to assume one, so we have not tested whether the

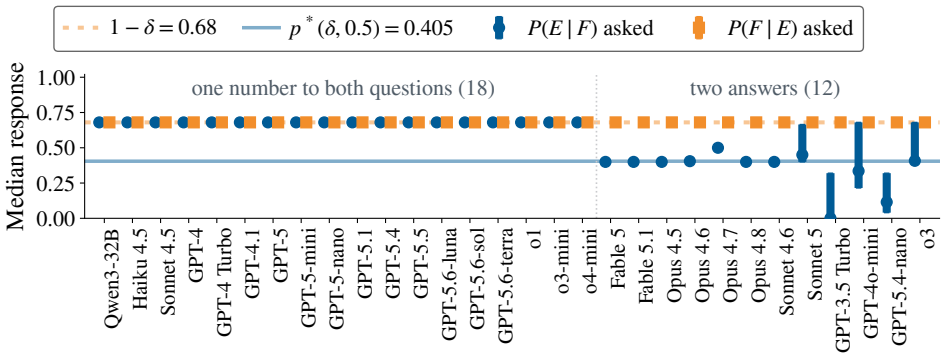


Figure 6: **Asked separately for the probability that the target exists and for the probability that the search missed it, 18 of 30 models give the same number to both.** Each model's median answer at $\delta = 0.32$, with each event described in words in its own one-turn call and no shared history. Most models show no interval because all their replies are identical. Dashed levels: the $1 - \delta$ rule at 0.68 and the posterior at a prior of 0.5, 0.405.

assumed prior changes with the domain. Therefore, a second call detects the substitution. Flagging a model whose two answers are within 0.02 of each other catches, among the 27 scored models with an assigned rule, 19 of the 20 on the $1-\delta$ rule and none of the seven on the posterior rule (App. C.1).

3.3 Practical downstream consequences of probability substitution. Pipelines that threshold the substituted model-reported probabilities can lead to unintended effects in downstream applications such as decision making, agent orchestration and multi-agent coordination. To demonstrate, we simulate a record creation pipeline that creates a record when the reported target probability is below 0.5. Thresholding can cause two types of errors: *suppression* (a false negative) when a record that does not exist is not created, and *duplicates* (a false positive) when a record exists and is wrongly created a second time. At a base rate of 0.01, thresholding the number, pooled over ten models, suppresses a record in 31.0% of decisions, against 0% for the correct posterior (Table 26). Which error type dominates depends on the base rate, with more suppressions for rare records at low base rates and more duplicates at high base rates (Fig. 7; Proposition 2).

Expected calibration error (Guo et al., 2017) scores the $1 - \delta$ responses as miscalibrated at any single base rate held fixed across coverages and as calibrated when the base rate equals $1/(1+\delta)$. Therefore, no calibration score can tell which event a report describes, and a calibration check passes the substituted number wherever the base rate equals $1/(1 + \delta)$ (Table 20).

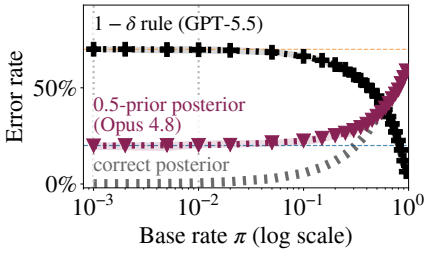


Figure 7: **At rare base rates, thresholding the substituted number loses records; the correct posterior does not.** Error rate at $\tau = 0.5$: GPT-5.5 ($1-\delta$ rule), Opus 4.8 (0.5 prior) and the correct posterior; dashed: each rule's ceiling.

Two standard repairs fit a rule to labeled outcomes: Platt scaling (Platt, 1999), a monotone map from the reported number to a probability, and conformal prediction (Mohri & Hashimoto, 2024), which withholds a decision when it cannot guarantee an error rate. Both repairs outperform the log-only posterior in most columns of Table 1. The conformal rule does so by sending most records to review. Platt scaling does so for a reason that does not transfer to deployment. The episodes include searches that found the record (about half of the present records overall, and most at the deeper coverage levels), where the model's number is high. On searches that did not find the record, the model's number does not contain any information beyond the coverage (App. H). Both repairs need labeled outcomes, which contain the base rate that a deployed pipeline lacks, and no function of the reported number alone gives the posterior at every base rate (Proposition 1, App. G).

The pipeline, not the model, should compute the posterior. The pipeline reads the coverage from the search log and the base rate from its own records, and applies Eq. 1 (the log-only rule of Table 1).

Table 1: **Percent of decisions wrong, missed plus duplicated records, at a threshold of** 0.5 **(lower is better, best method in bold).** The first five rows are ours (the model's number pooled over ten models, App. H), the three cited rows are published methods (Platt scaling is fitted to the true outcomes, and conformal prediction, at $\alpha = 0.10$, is refused where no threshold meets its guarantee), and the last two are reference policies; clinical 0.89 is that dataset's own base rate and pools all four coverage levels, the other base rates are set by us and never shown to the model, and the split is in Table 26; the stream's composition and the refits are stated in App. H.

% wrong	clinical 0.01	clinical 0.30	clinical 0.89	enterprise 0.01	enterprise 0.30
The model's number	31.2	28.7	25.8	29.2	29.1
Posterior, assumed rate	29.6	27.7	26.6	0.4	13.4
Posterior, true rate	0.5	18.3	27.4	0.6	18.4
Log only, assumed rate	49.8	43.0	41.7	0.6	18.6
Log only, true rate	0.6	18.6	21.7	0.6	18.6
Platt (Platt, 1999)	**0.4**	**11.8**	25.0	0.4	13.6
Conformal (Mohri & Hashimoto, 2024)	0.5	refused	refused	**0.2**	**1.4**
Margin (Chow, 1970)	29.5	25.1	**16.3**	27.7	25.2
Create every record	0.6	18.6	78.3	0.6	18.6
Review everything	0.0	0.0	0.0	0.0	0.0

At the clinical data's measured base rate of 0.89 that rule still makes errors, because 0.5 is the wrong threshold when records are common, not because the posterior is wrong (App. H). Where the base rate is unknown, the pipeline should record the posterior as undetermined rather than threshold the model's number. A two-call check detects a substituting model. Asked in two calls for the probability that the target exists and the probability that the search would have missed it, a model that gives both events the same number is flagged (Sec. 3.2).

4 Related Work

Absence, abstention and agent evaluations. Absence benchmarks vary how complete the evidence is and score categorical judgments. CROWN-QA (Min et al., 2026) documents over-closure from incomplete evidence, and Kim et al. (2026) find the coverage ratio over-predicted near zero. Abstention benchmarks measure when a model declines, including under penalties that make abstention optimal (Kirichenko et al., 2025; Wagner, 2026; Wang et al., 2026). Agent evaluations measure behavior under missing information and tools (Treviño et al., 2025; Li et al., 2025a; Zhao et al., 2026), browsing under uncertainty (Wei et al., 2025), and confidence after tool calls or during search (Xuan et al., 2026; Ou et al., 2026).

Verbalized confidence and its training. Verbalized confidence is scored on calibration (Lin et al., 2022; Kadavath et al., 2022) and acted on by conformal and abstention methods (Mohri & Hashimoto, 2024; Wen et al., 2025), and sampling-based detectors cluster a model's replies by meaning to score their spread (Kuhn et al., 2023; Farquhar et al., 2024). Confidence and abstention rewards are trained into models (Bani-Harouni et al., 2025; Stengel-Eskin et al., 2024; Xu et al., 2024; Zhang et al., 2024; Kapoor et al., 2024; Tan et al., 2026; Wei et al., 2026), alignment is reported to degrade calibration (Xiao et al., 2025; Leng et al., 2024), and Wagner & Abend (2026) ask what a model's probabilities can represent. We ask which event a reported probability refers to, where a fixed output schema is known to change what models report (Tam et al., 2024), and score each reply against the requested event.

Failed searches and base-rate reasoning. Search theory (Stone, 1975) and occupancy models in ecology and epidemiology (MacKenzie et al., 2002; Regan et al., 2006) compute the posterior that a target exists from the coverage of a search and a base rate. In human judgment, reporting $P(F \mid E)$ in place of $P(E \mid F)$ is the prosecutor's fallacy (Thompson & Schumann, 1987) and a form of base-rate neglect (Bar-Hillel, 1980; Casscells et al., 1978; Gigerenzer & Hoffrage, 1995; Oaksford & Hahn, 2004), and resembles attribute substitution (Kahneman & Frederick, 2002), while studies of language models find base-rate neglect and incoherent probability judgments (Saeedi et al., 2024; Zhu & Griffiths, 2024; Binz & Schulz, 2023; Scholten et al., 2024; Mondal et al., 2026; Imran et al.,

2025; Chen et al., 2026). In our setting the prompt omits the base rate, and most models identify the substituted event correctly when asked.

5 DISCUSSION

We find that about half of the LLMs we scored, 22 of 46, substitute the miss likelihood $1 - \delta$ for the posterior that the target exists. The substitution occurs when the prompt gives the coverage and withholds the base rate. A base rate in the prompt removes the substitution in most models, a probability field redefined as "the number you report" brings it back, and with no coverage at all in the prompt, most models give a probability below the base rate, as if the empty search showed that the target is absent (Secs. 3.1 and 3.2). Self-reported probabilities also pass between models in agent orchestration and multi-agent pipelines. When one model acts on another model's search report, the acting model's number and action depend on its own estimate, and a truthful "not found" leads the acting model to create a record that already exists (App. B).

We measure one calibration score, expected calibration error, and also test Platt scaling and a conformal rule (Sec. 3.3). Any method that uses only the reported number has the same limit. No function of that number gives the posterior at every base rate (Proposition 1). At one known base rate a monotone recalibration or a new threshold repairs the number; no single map works across base rates. Calibration checks, in forecasting (Murphy, 1973; Gneiting et al., 2007) and for agents reporting confidence after tool use (Kirchhof et al., 2025; Xuan et al., 2026; Ou et al., 2026; Lin et al., 2026), ask whether a number is calibrated, not what event the number describes. Methods that detect a missing premise (Fan et al., 2025; Li et al., 2025b; Kirichenko et al., 2025) or a number invented after a tool failure (Sethi et al., 2026; Kalai et al., 2026) do not catch the substitution either.

The model should not be asked for the posterior when the prompt withholds the base rate. Without the base rate the model cannot compute the posterior, and by default reports the miss likelihood instead. A field listing the inputs the reply is missing (Hou et al., 2024) leaves the substitution in place (Table 18).

Limitations. Our search tasks are controlled. The coverage is known by design and the reply is a JSON field, so free-form agents may behave differently. Where the model runs its own search, some models on the $1 - \delta$ rule keep the substitution and others drop it (Sec. 3.1), and we have not tested agentic benchmarks with free-form output fields. We classify responses by the rule they match and do not make any claim about what happens inside the model; the one positive activation-patching result (App. F) is our only evidence about a mechanism. 22 of the 46 models scored show the substitution, though 15 of the 22 are one provider's releases and by family the rule is the majority in 6 of 17 to 9 of 20 families, depending on the grouping (App. A.4). We measure the downstream harm only in a simulated pipeline on two datasets. Whether a model's written derivation reflects its computation is a separate question (Turpin et al., 2023; Lanham et al., 2023); our claims are based on the reported numbers, not on the written derivations.

Conclusion. A poorly chosen probability from a language model is usually read as a calibration failure. However, the LLMs we tested show a different failure. Given a partial search and no base rate, about half of the models fill the required probability field with the miss likelihood, a correct probability of the wrong event, and no recalibration of that number recovers the posterior at every base rate. The substitution is a default governed by the output field schema, not an inability, and no calibration score can tell which event the number describes, so pipelines that threshold the probabilities fail silently. Therefore, an evaluation should test which event a self-reported probability refers to before scoring how calibrated the probability is, and a pipeline should take the coverage from the search log and compute the posterior from its own base rate. The source of the default remains open. The base-and-instruct pairs we measured are consistent with the default arising before post-training and sharpening after, and fine-tuning three models of at most 14B parameters removed the miss-likelihood report in one of them but installed the posterior only where the prompt carries the training prompts' coverage line. Whether free-form agents and other underspecified probability questions show the same substitution is also open. The form of the request, not the model's knowledge, decides what a pipeline receives.

ETHICS STATEMENT

This work uses access-controlled, de-identified clinical datasets (MIMIC-III and MIMIC-IV) under their PhysioNet credentialed data use agreements. No patient-level content is redistributed; published materials contain derived aggregates and evaluation logs, and the raw model outputs the Reproducibility Statement provides use only the surrogate identifiers described below. Any original identifiers are replaced with surrogates, with automated checks verifying that every identifier was replaced. The clinical analyses are retrospective evaluations of model behavior and are not used to make or support clinical decisions. No human-subjects data were collected for this study.

REPRODUCIBILITY STATEMENT

Every case collected from September 2026 on had its design and scoring rule written into a numbered internal record before its first call (App. I), and every count the main text quotes from a case fixed in advance was checked against the raw model outputs by two independently written scorers. The benchmark generator, the scoring code, the raw model outputs, and the per-condition summaries are provided as supplementary material and will be made public. Replication details, seeds, controls, and assumptions are provided in App. I.

AI USE STATEMENT

In this work, we used generative AI tools for implementing the benchmark and evaluation code and for designing ablation experiments. We also used them for formulating and writing mathematical proofs, for feedback on experimental design, and for generating synthetic data sets. We used them for creating figures, drafting and editing the paper, and summarizing and finding related literature. We have not used generative AI tools for developing theoretical frameworks. We have reviewed all AI-assisted work and manually reproduced and verified the main results and findings. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv preprint arXiv:2503.02623*, 2025.

Maya Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, 1980.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Ward Casscells, Arno Schoenberger, and Thomas B. Graboys. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299(18):999–1001, 1978.

Chacha Chen, Matthew Jörke, Adam Goliński, Masha Fedzechkina, Guillermo Sapiro, Sinead Williamson, and Nicholas Foti. LLMs are not (consistently) Bayesian: Quantifying internal (in)consistencies of LLMs' probabilistic beliefs. *arXiv preprint arXiv:2605.06915*, 2026.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.06514.

Arsène Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

Albert Gong, Kamilė Stankevičiūtė, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, and Kilian Q. Weinberger. PhantomWiki: On-demand datasets for reasoning and retrieval evaluation. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Irving John Good. *Probability and the Weighing of Evidence*. Charles Griffin, London, 1950.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19023–19042, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Sohaib Imran, Ihor Kendiukhov, Matthew Broerman, Aditya Thomas, Riccardo Campanella, Rob Lamb, and Peter M. Atkinson. Are LLM belief updates consistent with Bayes' theorem? In *ICML 2025 Workshop on Assessing World Models*, 2025.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In Thomas Gilovich, Dale Griffin, and Daniel Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 49–81. Cambridge University Press, 2002.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051, 2026. Published version of arXiv:2509.04664, *Why Language Models Hallucinate*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2406.08391.

Remy Kim, Minseung Lee, Shuo Li, and Osbert Bastani. SeekerGym: A benchmark for reliable information seeking. *arXiv preprint arXiv:2604.17143*, 2026.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Position: Uncertainty quantification needs reassessment for large-language model agents. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2505.22655.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming Overconfidence in LLMs: Reward Calibration in RLHF. *arXiv preprint arXiv:2410.09724*, 2024.

Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

Belinda Z. Li, Been Kim, and Zi Wang. QuestBench: Can LLMs Ask the Right Question to Acquire Information in Reasoning Tasks? In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025a. arXiv:2503.22674.

Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. Don't take the premise for granted: Evaluating the premise critique ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025b. arXiv:2505.23715.

Moule Lin, Qizhen Lan, Shuhao Guan, Weipeng Jing, Jiexin Fan, David Gregg, and Goetz Botterweck. Uncertainty quantification for LLM agents: A taxonomy, an evaluation protocol, and an empirical study. *arXiv preprint arXiv:2609.07395*, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Darryl I. MacKenzie, James D. Nichols, Gideon B. Lachman, Sam Droege, J. Andrew Royle, and Catherine A. Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

Michelle McDowell and Perke Jacobs. Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12):1273–1312, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Byoungjae Min, Kennedy Edemacu, Sae-Hong Cho, Yoonhyuk Choi, Beakcheol Jang, and Jong Wook Kim. When absence is evidence: Evaluating completeness-sensitive negative reasoning in large language models. *arXiv preprint arXiv:2608.04591*, 2026.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.10978.

Manuel Mondal, Ljiljana Dolamic, Gérôme Bovet, Philippe Cudré-Mauroux, and Julien Audiffren. Implicit Probabilistic Reasoning Does Not Reflect Explicit Answers in Large Language Models. *Transactions on Machine Learning Research*, 2026. arXiv:2406.14986.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Jay I. Myung and Mark A. Pitt. Optimal experimental design for model discrimination. *Psychological Review*, 116(3):499–518, 2009.

Mike Oaksford and Ulrike Hahn. A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58(2):75–85, 2004.

Litu Ou, Kuan Li, Huifeng Yin, Liwen Zhang, Zhongwang Zhang, Xixi Wu, Rui Ye, Zile Qiao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. BrowseConf: Confidence-guided test-time scaling for web agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 446–465, 2026.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

Tracey J. Regan, Michael A. McCarthy, Peter W. J. Baxter, F. Dane Panetta, and Hugh P. Possingham. Optimal eradication: When to stop looking for an invasive plant. *Ecology Letters*, 9(7): 759–766, 2006.

Payam Saeedi, Mahsa Goodarzi, and M Abdullah Canbaz. Heuristics and Biases in AI Decision-Making: Implications for Responsible AGI. *arXiv preprint arXiv:2410.02820*, 2024.

Florian Scholten, Tobias R. Rebholz, and Mandy Hütter. Metacognitive Myopia in Large Language Models. *arXiv preprint arXiv:2408.05568*, 2024.

Arham Sethi, Arsen Kenzhebayev, Saanvi Paturi, Vatsal Raina, Vyas Raina, and Ivaxi Sheth. Fabrication after tool failure: Tool-augmented agents assert values their tools did not return. *arXiv preprint arXiv:2609.14758*, 2026.

Elliott Sober. Absence of evidence and evidence of absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies*, 143(1):63–90, 2009.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2405.21028.

Christopher Stephens. A Bayesian approach to absent evidence reasoning. *Informal Logic*, 31(1): 56–65, 2011.

Lawrence D. Stone. *Theory of Optimal Search*. Academic Press, New York, 1975.

Yuhong Sun, Joachim Rahmfeld, Chris Weaver, Weijia Chen, Roshan Desai, Wenxi Huang, and Mark H. Butler. EnterpriseRAG-Bench: A RAG benchmark for company internal knowledge. *arXiv preprint arXiv:2605.05253*, 2026.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 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*, pp. 1218–1236, 2024. doi: 10.18653/v1/2024.emnlp-industry.91. arXiv:2408.02442.

Chee Heng Tan, Zhuoyi Lin, Mehul Motani, and Wee Sun Lee. On the effectiveness of reward functions in reinforcement learning for confidence calibration of large language models. *arXiv preprint arXiv:2607.04332*, 2026.

William C. Thompson and Edward L. Schumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11 (3):167–187, 1987.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Eduardo Treviño, Hugo Contant, James Ngai, Graham Neubig, and Zora Zhiruo Wang. Benchmarking failures in tool-augmented language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.04388.

Benedikt J. Wagner. Two axes of LLM abstention: Answer correctness and question answerability. *arXiv preprint arXiv:2607.08456*, 2026.

Eitan Wagner and Omri Abend. Express your doubts: Probabilistic world modeling should not be based on token logprobs. In *International Conference on Machine Learning (ICML), Position Paper Track*, 2026. arXiv:2505.02072; v1 (2025) titled "What do Language Model Probabilities Represent? From Distribution Estimation to Response Prediction".

Jiawei Wang, Yanfei Zhou, Siddartha Devic, and Deqing Fu. Are LLM decisions faithful to verbal confidence? *arXiv preprint arXiv:2601.07767*, 2026.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 2025. arXiv:2407.18418.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. Restoring Calibration for Aligned Large Language Models: A Calibration-Aware Fine-Tuning Approach. In *International Conference on Machine Learning*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2405.20974.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Yunze Xiao, Junjue Wang, and Naoto Yokoya. The confidence dichotomy: Analyzing and mitigating miscalibration in tool-use agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921, 2018.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, 2024. arXiv:2311.09677.

Jiale Zhao, Ke Fang, and Lu Cheng. When and What to Ask: AskBench and Rubric-Guided RLVR for LLM Clarification. *arXiv preprint arXiv:2602.11199*, 2026.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent Probability Judgments in Large Language Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2024.

A THE BENCHMARK

This part describes the NullSearch benchmark and how each model is assigned a response rule: the environment, whose coverage δ is a design constant of every episode; the conditions and reply forms; the 60% criterion behind the 22 / 7 / 17 split; the roster of the 46 releases (Table 7); and every classified model's curve (Fig. 9).

A.1 THE GENERATED ENVIRONMENT

Every record is planted by code and the environment is small enough to list whole in a prompt, so a failed search has really failed and its coverage is known exactly.

Design. The environment is a code-generated fictional organization storing typed records across 6 systems $\times$ 3 sections $=$ 18 locations. Each episode asks about one record of one *record type*, and each record type has a six-location *candidate set*, the only places a record of that type is ever stored, with location shares summing to one, which the prompts state. The share table is fixed in the design, not redrawn per episode, so each episode's δ is a known design constant and coverage levels are comparable across episodes. A record the generator did not plant cannot exist, so absence is confirmed by construction. The coverage of a set of searches is the summed location share of the searched candidate locations; searching a non-candidate section contributes zero. The zero-coverage level covers six non-candidate sections, so the prompt shows visible search effort although $\delta = 0$; the *unsearched* condition, by contrast, shows no search log. The $1-\delta$ rule and both associated behaviors are unchanged at 102, at 1,000 and at 10,000 locations and under permuted shares (App. E.2). The nine transfer settings of App. B reuse the same measurement on external data.

Coverage levels. Level 0 is the zero-coverage search, levels 1–4 are partial coverage searches, and level 5 is exhaustive search coverage; Table 2 gives the δ range at each level.

Table 2: **The six coverage levels run from a guaranteed empty search to confirmed absence, and every partial level except level 2 tells the three reference rules apart at the 0.02 tolerance.** For each level, the δ range, searched candidates, and role.

Level	δ (range seen)	Searched candidates	Note
0	0	none (6 non-candidates)	zero coverage, guaranteed empty a priori
1	0.12–0.13	two smallest-share locations	separates all three reference rules
2	0.31–0.36	e.g. 22% + 10%	mid-coverage level
3	0.51–0.57	e.g. 40% + 16%	mid-high coverage level
4	0.76–0.78	three locations	high-coverage level
5	1.0	all six	confirmed absence of target

A.2 EPISODES AND CONDITIONS

An *episode* pairs one question about one search task ("does a parking_permit record exist for Omar Okafor?") with a set of searches, and the conditions change the searches or their summary while the question stays fixed. The prompt is templated with the question, inventory, shares, and a search log listing each search with its result. In the *absent* condition the record exists nowhere; in the *present* condition it exists, and the search either finds it (the log shows a `FOUND` line) or misses it.

Two coverage variations differ by a single line: the coverage-stated variation appends a summary stating the searched total, and the coverage-withheld variation omits it, so the model must pick the searched candidates out of a log that also lists non-candidate sections and sum their shares itself. A third variation withholds the shares and shows only the candidate list, so the model can only count (App. F.2). The unsearched condition removes the search log entirely and elicits the effective prior.

Response parsing. The last JSON object in the response must contain a valid probability number; anything else is counted as unparsed, never defaulted or fixed, and excluded from metrics. A run in

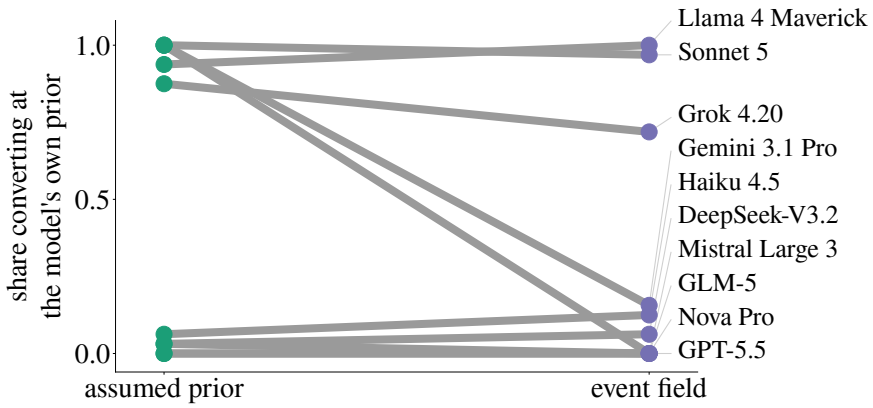
which a model exceeds the 25% unparsed-response allowance is excluded from analysis entirely, so every count in this part runs over the models under that allowance.

A.3 THE REPLY FORMS

The probability field is the primary elicitation, a one-turn JSON reply asking for the posterior $P(E \mid F)$ that the record exists. The other forms change what the reply asks for while the search stays fixed, so each tests which event the number describes.

- *Event field (one turn).* A field beside the number asks which event the probability describes, in the same reply.
- *Event question (two turns).* A second turn, after the number is fixed, asks which event the earlier probability describes.
- *Prior field.* The prompt either states a base rate, as a number or as a background fact, or requires the model to give the prior it assumed beside its number.
- *Both.* One form adds the event field and the prior field together.
- *Decline.* A decline option lets the model return that the probability is not determinable.

The event-field form and the form with both fields redefine the probability field as "the number you report"; the prior-field form keeps the existence wording. Table 17 (App. D.4) gives each one-turn form's definition in the templates' words, with the matched form of App. D.3 and the prefilled prior. Fig. 8 shows how often ten models reach the posterior under each form.



not drawn: Kimi K2.7

Figure 8: **An assumed prior converts half of the ten, and requiring the model to select the event reverses that in two of those five, weakening a third.** How often each reaches the posterior in each ask, over the ten drawn. Kimi K2.7 is not drawn: 9 of its 32 combined replies do not parse, over the 25% allowance.

A.4 ASSIGNING A RULE TO EACH MODEL

A model is assigned a rule by the share of its partial-coverage responses on that rule's curve, and the assignment holds when the threshold, the resampling unit or a free prior per model is changed.

The criterion. Three reference rules are candidates: the $1 - \delta$ rule, the 0.5-prior posterior and the counting rule. A model is assigned a rule when at least 60% of its polled responses at the four partial-coverage levels fall within 0.02 of that rule's curve, over 122 to 1,478 responses per model. A response inside two candidates' tolerances distinguishes nothing and is out of the vote, and a model for which no rule reaches 60% is called mixed. The endpoints do not enter: at $\delta = 0$ and $\delta = 1$ the candidate rules predict values closer than the 0.02 range, and at level 2 $1 - \delta$ and the counting rule fall within 0.007 to 0.017 of each other on 51% of episodes, so there the within-level slope decides without a tolerance (below). Under this criterion the 46 models split 22 / 7 / 17 between the $1 - \delta$ rule, the posterior and the mixed group (Table 7).

Changing the threshold. Table 3 re-assigns all 46 releases at thresholds from 50 to 75%. None changes from one rule to the other; every move is from a given rule into the mixed rule group.

Table 3: **Raising the criterion from 60% to 70% moves 5 of 46 releases, all into the mixed rule group; none changes rule.** Counts of $1-\delta$ / posterior / mixed rule assignments at each threshold (top), and the rows whose assignment differs between 60% and 70% (bottom) with the assigned-rule share at 60% and the Wilson 95% interval on that share over its n polled responses. At 60%, 4 assigned releases have intervals that include the boundary.

Threshold	$1-\delta$ rule	posterior	mixed rule
50%	25	7	14
55%	23	7	16
60%	22	7	17
65%	19	6	21
70%	18	6	22
75%	17	6	23

Release (provider)	Rule at 60%	Share	n	Wilson 95%
Sonnet 5 (Anthropic)	$1-\delta$ rule	65.5%	574	[61.5%, 69.3%]
GPT-5.1 (OpenAI)	$1-\delta$ rule	61.7%	256	[55.6%, 67.5%]
GPT-4 (OpenAI)	$1-\delta$ rule	60.9%	128	[52.3%, 69.0%]
GPT-4o (OpenAI)	$1-\delta$ rule	60.9%	128	[52.3%, 69.0%]
Llama 4 Maverick (Meta)	posterior	60.2%	128	[51.5%, 68.2%]

Resampling by family. The survey's 22 of 46 has the exact 95% interval $[33, 63]$. Resampling families, not models (2,000 bootstrap samples), gives $[29, 63]$, $[31, 63]$ or $[33, 63]$ with the three GPT-5.6 releases inside their generation's family (17 families), as one of their own (18) or as three (20), and the rule is the majority in six, seven or nine of those families (Table 4).

One free prior per model. A single free prior fitted to each release's four partial-level medians places 13 of the 17 mixed-rule models closer to a fixed-prior posterior than to $1-\delta$, and the remaining four closer to $1-\delta$ (Table 5). Scored at the response level with the prior fitted per model and each level kept out in turn, none of the 17 passes.

Table 4: **Resampling by family widens the interval in all but one grouping.** 22 of the 46 releases report the miss likelihood; over releases the interval is exact, 33 to 63%. By family, the rule is the majority in 6 of 17, 7 of 18 or 9 of 20, and resampling widens the interval to 29 to 63%, 31 to 63% and 33 to 63%; the row shows the widest.

Models	$1-\delta$	95% interval (%)
17 families	6 of 17	29–63

A.4.1 SLOPE IDENTIFICATION

The within-level slope separates the counting rule from $1-\delta$ with no tolerance. Within a level the counting rule is a constant, since the number of searched candidates is fixed, while $1-\delta$ varies with the searched locations' shares, so the slope is -1 for the $1-\delta$ rule, 0 for any constant, and $-1/(2-\delta)^2 \approx -0.36$ for the 0.5-prior posterior. Over the second level's four coverage values GPT-5.5's most common response falls one-for-one with δ and Opus 4.8's at the posterior's rate, whereas the counting rule, constant at 0.667 within the level, fits neither. Over the fifty-four configurations with usable within-level slopes, including the open-weight scale series, 27 of the 54 show the rule's slope of -1.00 at each level tested when the coverage total is in the prompt, and 16 of 54 when it is withheld and the shares must be summed (Table 6). A rule outside the candidates that equals $1-\delta$ at every tested coverage would remain indistinguishable, so the labels claim identification within the listed candidates only.

Table 5: A single free prior fitted to each release's four partial-level medians. RMS is the root-mean-square gap between the medians and the named curve, so lower is better, as the arrows mark. The best of the three is in bold. Of the 17 mixed rule models, 13 are closer to a free-prior posterior than to $1-\delta$, and 8 of the 17 fitted priors lie in $[0.4, 0.6]$.

Release	Rule	$\hat{\pi}$	RMS($\hat{\pi}$) $\downarrow$	RMS($1-\delta$) $\downarrow$	RMS(0.5) $\downarrow$	Best
DeepSeek-V3.2	mixed	0.57	**0.015**	0.196	0.068	free prior
Gemma-3-27B-it	mixed	0.62	**0.076**	0.145	0.140	free prior
GLM-5	mixed	0.68	0.138	**0.057**	0.226	$1-\delta$
GPT-3.5 Turbo	mixed	0.39	**0.149**	0.407	0.174	free prior
GPT-4.1	mixed	0.56	0.216	**0.214**	0.222	$1-\delta$
GPT-4.1-mini	mixed	0.59	**0.021**	0.177	0.089	free prior
GPT-4.1-nano	mixed	0.01	**0.006**	0.607	0.358	free prior
GPT-4o-mini	mixed	0.19	**0.036**	0.485	0.243	free prior
GPT-5.2	mixed	0.63	0.159	**0.118**	0.206	$1-\delta$
GPT-5.4-mini	mixed	0.55	**0.189**	0.209	0.194	free prior
GPT-5.4-nano	mixed	0.06	**0.008**	0.571	0.319	free prior
GPT-OSS-120B	mixed	0.52	**0.013**	0.248	0.019	free prior
Llama-3.3-70B-Instruct	mixed	0.69	0.138	**0.028**	0.244	$1-\delta$
Mistral Large 3	mixed	0.33	**0.089**	0.407	0.167	free prior
o3	mixed	0.49	**0.019**	0.262	0.019	free prior
Opus 4.5	mixed	0.49	**0.007**	0.269	0.012	free prior
Opus 4.7	mixed	0.58	**0.038**	0.205	0.084	free prior

Table 6: **Observed within-level slopes sort the configurations into the same groups as the rule assignments, with no tolerance involved.** Within a coverage level the counting rule has slope 0, the $1-\delta$ rule slope -1, and the 0.5-prior posterior slope near -0.36. Denominators are the slope test's 54 configurations, labeled by this test's own nearest-value criterion, not the classification of Table 7.

Behavior	Predicted slope	Observed	Configurations
$1-\delta$ rule	-1	-1.00 exactly	27/54 stated; 16/54 withheld
0.5-prior posterior	$-1/(2-\delta)^2 \approx -0.36$	-0.38 to -0.46	2 models
constants	0	0.00	see text
mixed	none stable	sign reverses across levels	3 models

A.5 THE MODELS MEASURED

Table 7 lists all 46 releases with the rule assigned, its share and the evaluation size, sorted by rule; the largest group follows the $1 - \delta$ rule, 22 of 46, with 17 mixed rule models and seven posterior models.

Rule assignment takes 25 OpenAI, ten Anthropic, and 11 releases from eight further providers. Most follow-up experiments use GPT-5.5 and Opus 4.8, the most consistent example of each rule. The two newest releases, GPT-5.6-sol and Fable 5.1, are shown in the main figures in every case measured on the same episodes. Kimi-K2.6 returned no parsed response in any of its 256 queries and is unclassified, and a few API-hosted releases could not be measured. Opus 5 refuses this instrument on the direct channel and gave no parseable response on the second channel, so it is in no count; the record is in App. I.

A.6 PER-MODEL CURVES

Every classified model's response curve is drawn in Figs. 9 and 10, one part per release, grouped by the rule the criterion assigns. Kimi-K2.6 is unclassified (App. A.5) and not drawn; six models have responses at the four partial-coverage levels only and so no full-coverage point. GPT-5.4-nano lies below each reference rule and is a mixed rule model under the criterion (Sec. 2).

B BEYOND THE BENCHMARK

This part reports the substitution outside the generated benchmark: on external datasets, when the model runs the search itself, when the coverage arrives in another form or not at all, and when the field asks for an amount or a count instead of a probability. The main text's claims that the default survives a search the model runs (Sec. 3.1), that it is specific to a probability field over a stated coverage, and that the benchmark was applied to nine transfer settings rest on the results here. Across the settings, what a model reports is fixed by the most precise coverage the setting lets it work out.

B.1 THE DATASETS AND WHAT RAN ON EACH

Every setting confirms absence by construction and defines the detection probability in its own way, so no ground truth depends on labeling. The search is a record in the prompt in every setting but three: on EnterpriseRAGBench and in the code search the model chooses its searches and δ is read off its log, and PhantomWiki's retrieval mode shows the top-k articles a vector search surfaced, none mentioning the subject.

MIMIC-III and MIMIC-IV contribute de-identified clinical records: ICU, hospital medication, emergency-department, and notes-crossed-with-emergency-department tables (Johnson et al., 2016; 2023). Spider-ACL is an access-control derivative of the Spider text-to-SQL databases whose queries execute live (Yu et al., 2018). DDXPlus provides synthetic patient intake records that are complete by construction (Fansi Tchango et al., 2022), and PhantomWiki a generated encyclopedia whose articles enumerate every family relation (Gong et al., 2025). In the last two settings the agent runs the search itself: EnterpriseRAGBench, a 511,962-document corpus searched under a checking budget (Sun et al., 2026), and a generated code library searched by live grep. The nine are each MIMIC table, Spider-ACL, DDXPlus, PhantomWiki, EnterpriseRAGBench and the code search.

What ran on each:

- **MIMIC-III and MIMIC-IV** (four settings): the designed-miss condition with coverage stated and withheld (App. B.2).
- **Spider-ACL**: live queries, some denied; GPT-5.5 treats a denied query as evidence of presence.
- **DDXPlus**: twinned endpoints, a demographics-only record against the full intake (App. B.6).
- **PhantomWiki**: twinned endpoints, unrelated articles against the subject's own article, and an even share table crossed with the coverage line present or dropped (App. B.6).
- **EnterpriseRAGBench**: the model chooses which sections to check, once with stated shares and once told its coverage as a running percentage (Apps. B.3 and B.4).

Table 7: **All 46 releases under one criterion: 22 follow the $1-\delta$ rule, 7 the posterior, 17 mix rules.** A rule is assigned at $\geq 60\%$ of polled responses at the four partial-coverage levels, within 0.02; responses inside both tolerances are out of the vote. Match is the assigned rule's share; mixed rule rows read $1-\delta$ share / posterior share; n counts polled responses. Five of the seven posterior releases are Anthropic's. Llama 4 Maverick passes the criterion by one response.

Release (provider)	Rule	Match	n
Nova Pro (Amazon)	$1-\delta$ rule	72.4%	127
Haiku 4.5 (Anthropic)	$1-\delta$ rule	96.1%	256
Sonnet 4.5 (Anthropic)	$1-\delta$ rule	82.8%	128
Sonnet 5 (Anthropic)	$1-\delta$ rule	65.5%	574
DeepSeek-R1 (DeepSeek)	$1-\delta$ rule	100.0%	256
Kimi K2.7 (Moonshot)	$1-\delta$ rule	85.1%	235
GPT-4 (OpenAI)	$1-\delta$ rule	60.9%	128
GPT-4 Turbo (OpenAI)	$1-\delta$ rule	77.3%	128
GPT-4o (OpenAI)	$1-\delta$ rule	60.9%	128
GPT-5 (OpenAI)	$1-\delta$ rule	100.0%	241
GPT-5-mini (OpenAI)	$1-\delta$ rule	97.7%	128
GPT-5-nano (OpenAI)	$1-\delta$ rule	81.2%	128
GPT-5.1 (OpenAI)	$1-\delta$ rule	61.7%	256
GPT-5.4 (OpenAI)	$1-\delta$ rule	93.8%	576
GPT-5.5 (OpenAI)	$1-\delta$ rule	100.0%	1478
GPT-5.6-luna (OpenAI)	$1-\delta$ rule	98.4%	128
GPT-5.6-sol (OpenAI)	$1-\delta$ rule	100.0%	128
GPT-5.6-terra (OpenAI)	$1-\delta$ rule	99.2%	128
o1 (OpenAI)	$1-\delta$ rule	100.0%	128
o3-mini (OpenAI)	$1-\delta$ rule	98.4%	128
o4-mini (OpenAI)	$1-\delta$ rule	100.0%	128
Grok 4.20 Reasoning (xAI)	$1-\delta$ rule	100.0%	128
Fable 5 (Anthropic)	posterior	100.0%	128
Fable 5.1 (Anthropic)	posterior	100.0%	128
Opus 4.6 (Anthropic)	posterior	93.0%	128
Opus 4.8 (Anthropic)	posterior	98.6%	1101
Sonnet 4.6 (Anthropic)	posterior	78.1%	128
Gemini 3.1 Pro (Google)	posterior	91.0%	122
Llama 4 Maverick (Meta)	posterior	60.2%	128
Opus 4.5 (Anthropic)	mixed rule	7.8% / 49.6%	528
Opus 4.7 (Anthropic)	mixed rule	50.0% / 17.2%	128
DeepSeek-V3.2 (DeepSeek)	mixed rule	28.1% / 5.1%	512
Gemma-3-27B-it (Google)	mixed rule	37.5% / 4.3%	256
Llama-3.3-70B-Instruct (Meta)	mixed rule	59.0% / 0.0%	256
Mistral Large 3 (Mistral)	mixed rule	10.2% / 7.0%	128
GPT-3.5 Turbo (OpenAI)	mixed rule	26.6% / 4.8%	124
GPT-4.1 (OpenAI)	mixed rule	29.7% / 3.9%	128
GPT-4.1-mini (OpenAI)	mixed rule	47.7% / 5.5%	128
GPT-4.1-nano (OpenAI)	mixed rule	3.1% / 0.0%	128
GPT-4o-mini (OpenAI)	mixed rule	24.2% / 0.8%	128
GPT-5.2 (OpenAI)	mixed rule	43.0% / 3.9%	128
GPT-5.4-mini (OpenAI)	mixed rule	4.7% / 0.8%	128
GPT-5.4-nano (OpenAI)	mixed rule	16.9% / 1.1%	272
o3 (OpenAI)	mixed rule	35.2% / 25.0%	128
GPT-OSS-120B (OpenAI, open-weight)	mixed rule	50.4% / 46.9%	256
GLM-5 (Zhipu)	mixed rule	22.7% / 3.9%	128

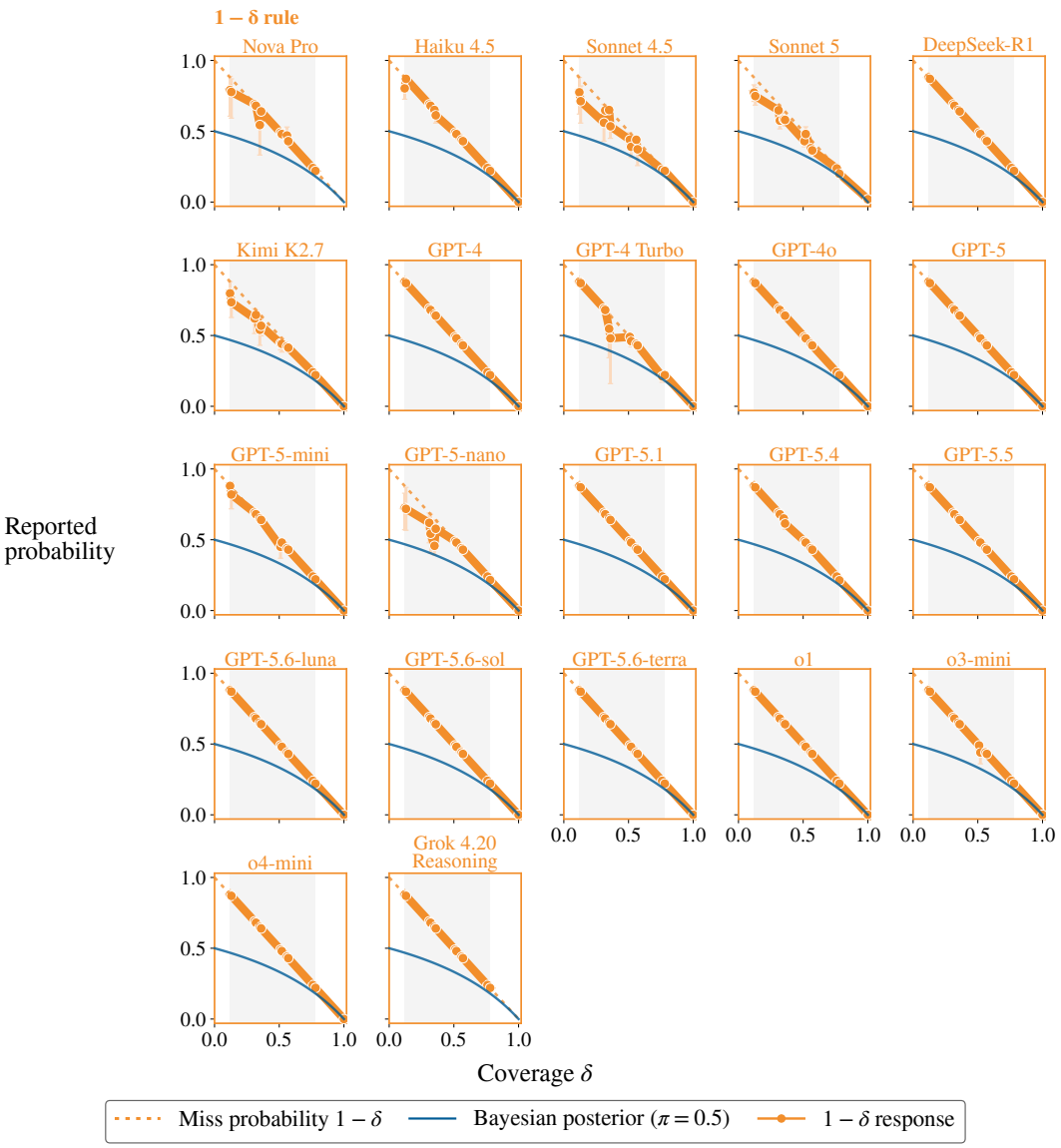


Figure 9: **22 models trace the $1-\delta$ line, 7 the posterior, and 17 are mixed at the criterion.** Per-model response curves, grouped and colored by assigned rule under the 60% match-rate criterion: one part per classified release (46 parts), named and ordered as in Table 7. The upper parts hold the $1-\delta$ group; Fig. 10 below continues with the posterior and mixed rule groups. Markers are per-level means with bootstrap intervals; the shaded band marks the four partial-coverage levels. Conditions as in Fig. 2, whose markers are per-level medians.

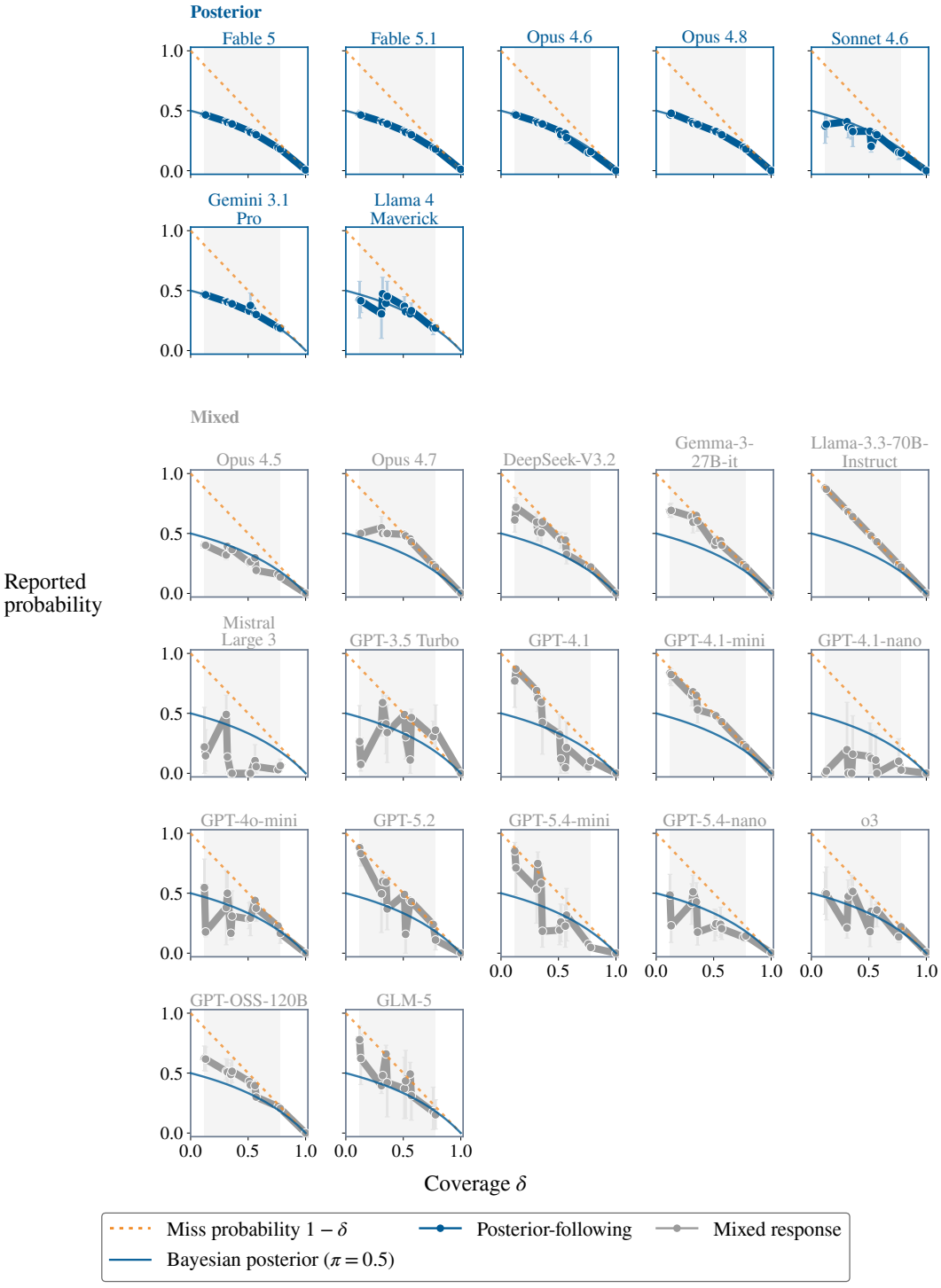


Figure 10: **Per-model response curves (continued).** The 7 posterior and 17 mixed rule models of Fig. 9; markers, band, and conditions as there.

- **The generated code library**: the live grep search, and a delegated variant in which a second model reads the searcher's report (App. B.5).
- **The software-testing programs** (a test run rather than a search, outside the nine): the bug form (App. B.7).

The clinical records and the enterprise documents also supply the episodes of the deployment simulation (Sec. 3.3; App. H). The harnesses, prompts and per-model tables for all nine settings are issued with the paper (App. I).

B.2 MIMIC: CLINICAL RECORDS

On both hospital medication tables, GPT-5.5's most common response after a miss is that table's own $1 - \delta$ (Table 8).

PhysioNet's de-identified MIMIC-III and MIMIC-IV databases, accessed under their data use agreements, contribute four settings: ICU records, hospital medication records, emergency-department records, and MIMIC-IV notes crossed with the emergency-department tables. A patient is absent for a record category exactly when the subject id appears in the survey table but in none of the candidate tables. Location shares and base rates are measured from the data: the emergency-department medication tables (77.4% medrecon and 22.6% pyxis) at a base rate of 0.89, and the hospital medication tables of Table 8 at 0.72. In the notes setting only, δ is estimated from a term list rather than measured. Nothing from these databases is copied into the repository.

A designed-miss condition assigns the searched table so that every episode misses: with coverage stated, GPT-5.5 matches $1 - \delta$ in most large-table episodes, and with coverage withheld no response matches it (0 of the 64 episodes).

Table 8: **On both MIMIC medication tables GPT-5.5's most common miss response is that table's own $1 - \delta$.** GPT-5.5 on prescriptions and emar (the electronic medication-administration record), measured base rate 0.72. Rates count the most common response when the searched table is assigned so each episode misses.

Hospital table	Location share	Posterior after a miss	Most common response	Rate
prescriptions	75.3%	0.384	0.25 $(1 - \delta)$	20 of 32
emar	24.7%	0.656	0.75 $(1 - \delta)$	11 of 32

B.3 ENTERPRISE RAG: THE MODEL CHOOSES WHAT TO CHECK

A model that chooses which sections of enterprise documents to check reports the miss likelihood of the stated shares in almost every episode: GPT-5.5 does so in 63 of 64, after examining at most 8% of the documents.

The 511,962-document corpus is used as issued. Our wrapper adds the location shares in the prompt, as in the benchmark, a bounded checking procedure, and a log from which the achieved δ is computed; absence is confirmed by the wrapper's deterministic generator. GPT-5.5 treats the stated inventory as complete instead of deriving coverage from its own log, and when the two conflict it follows the inflated value. The pipeline downstream then receives the share of the corpus the search never read, a number with no bearing on whether the record exists.

B.4 COVERAGE AS A RUNNING PERCENTAGE

Told its coverage as a percentage after every scan, and given no shares, no model reports the miss likelihood in more than one episode of its 16.

In this second case on the same corpus, 11 models each ran 16 absent episodes at budget 8 and stopped at a median coverage of 13%. One model reports the miss likelihood in 1 of 16 episodes and the others in none. Four of the 11 report 0.02 or less in at least 14 of 16 episodes, two more do so in 6 and 3, four give medians between 0.10 and 0.18, and one's replies are malformed in 14 of 16. The two cases differ only in how the coverage reaches the model.

B.5 THE LIVE GREP SEARCH

When the model runs the search itself, two of the five readable releases on the $1-\delta$ rule keep reporting the miss likelihood of the search they ran, and three report less (Table 9; Fig. 11).

The setting asks "does a function named X exist in this repository?" of a generated Python library, through a `grep` tool under a query budget B; the model writes a JSON object that the harness reads and runs. Within a budget the achieved coverage does not vary, so only the move between budgets is read. A release whose unparsed share passes the 10% limit set in advance is shown and not classified.

Six releases on the $1-\delta$ rule have run it. One, Kimi K2.7, cannot be read: 35 and 36 of its 40 replies do not parse at the two budgets. Of the five that can, GPT-5.5 and Grok 4.20 Reasoning carry the rule into the live search at gaps of 0.000, and Haiku 4.5, Nova Pro and Sonnet 5 depart from it. Table 9 gives each release and budget except Sonnet 5, which is measured by a move over 40 targets matched across two budgets: it shifts -0.081, where $1-\delta$ would shift -0.160 and the posterior -0.095, rejecting $1-\delta$ at 9.2 standard errors and sitting 1.6 from the posterior.

Table 9: **The live grep search, per release and budget.** Median stated probability against 1 minus the median achieved coverage, at each budget, with the count that parsed. A release over the 10% unparsed limit set in advance is shown and not classified. Within a budget the achieved coverage does not vary, so only the move between budgets is read here. GPT-5.5's budgets 2 and 3 come from an earlier run of the same harness, and every other row from a single later run. A row matches when its median is within 0.05 of $1-\delta$, the tolerance set in advance for this run, wider than the 0.02 used elsewhere; Gemini 3.1 Pro's medians, 0.275 and 0.180, equal the posterior at a prior of 0.5 at both budgets.

Release	Rule	B	Scored	Unparsed	Median vs $1-\delta$	Gap
GPT-5.5	$1-\delta$	1	40	0	0.600 vs 0.600	0.000 matches
GPT-5.5	$1-\delta$	2	40	0	0.380 vs 0.380	0.000 matches
GPT-5.5	$1-\delta$	3	40	0	0.220 vs 0.220	0.000 matches
GPT-5.5	$1-\delta$	4	40	0	0.120 vs 0.120	0.000 matches
Grok 4.20 Reasoning	$1-\delta$	2	40	0	0.380 vs 0.380	0.000 matches
Grok 4.20 Reasoning	$1-\delta$	3	40	0	0.220 vs 0.220	0.000 matches
Haiku 4.5	$1-\delta$	2	40	0	0.270 vs 0.380	0.110 departs
Haiku 4.5	$1-\delta$	3	40	0	0.150 vs 0.220	0.070 departs
Kimi K2.7	$1-\delta$	2	5	35 of 40	–	not classified
Kimi K2.7	$1-\delta$	3	4	36 of 40	–	not classified
Nova Pro	$1-\delta$	2	40	0	0.150 vs 0.380	0.230 departs
Nova Pro	$1-\delta$	3	40	0	0.050 vs 0.220	0.170 departs
Gemini 3.1 Pro	posterior	2	40	0	0.275 vs 0.380	0.105 departs
Gemini 3.1 Pro	posterior	3	38	2	0.180 vs 0.220	0.040 matches
Opus 4.8	posterior	2	40	0	0.280 vs 0.380	0.100 departs
Opus 4.8	posterior	3	40	0	0.120 vs 0.220	0.100 departs

Gemini 3.1 Pro: the two budgets fall on opposite sides of the tolerance, so no single verdict is given for the release.

Delegated search. When the same evidence arrives as another model's report, the number changes and then passes straight into the action. A worker model runs the grep search above and reports what it found; an orchestrator model reads only that report and states the posterior or commits to an action under given costs, in three model families. Opus 4.8, near-certain from its own search log, answers a median 0.50 when the identical evidence arrives as a worker's report. Two models in different families act on their own belief in 240 of 240 committed decisions and never escalate,

and where the worker's truthful report says "not found" while the target exists, both create (48 of 48 each). A decision output is therefore a fix only where it is separately validated (Sec. 3.3).

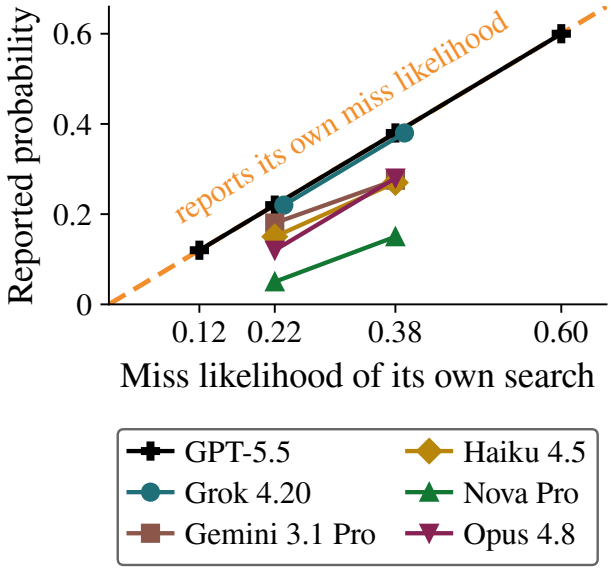


Figure 11: **The substitution survives when the model runs the search itself: two models report the miss likelihood of their own search at every budget, and the rest report less.** Each line is one model: x is the miss likelihood of its own search at each grep budget, one minus the median coverage its searches achieved, and y its median reported probability over 40 episodes, over the replies that parse, the per-model summary used in Fig. 2. On the dashed diagonal a model reports its own miss likelihood; GPT-5.5 and Grok 4.20 Reasoning lie on it, and Gemini 3.1 Pro, Haiku 4.5, Opus 4.8 and Nova Pro below. At a prior of 0.5 the posterior (Eq. 1) for a miss likelihood of 0.60, 0.38, 0.22 and 0.12 is 0.375, 0.275, 0.180 and 0.107. Gemini 3.1 Pro's medians, 0.275 and 0.180, equal the 0.5-prior posterior at both budgets; the other three below the diagonal have no tested account.

B.6 No coverage: the empty search

Given no coverage of any kind, models treat an empty result as absence: five of the eight models tested report below the range of the base rate.

Zero-coverage answers are deterministic per prompt and depend on the prompt's wording and line order. Reordering only the search-log lines moves about half the deterministic prompts to a different response (23 of 40 coverage-withheld, 25 of 53 coverage-stated), whereas an open-weight model that works its answers out from the shares is untouched (0 of 26).

Two external datasets set an empty endpoint against a complete one on the same episodes, for thirteen releases. On DDXPlus, a demographics-only record against the full intake: where nothing was searched the models answer with the population base rate, and four releases answer lower at the full intake in every twinned episode. On PhantomWiki, unrelated articles against the subject's own article: 11 releases separate the two endpoints, one did not, and one shifts without crossing the confirmed-response threshold.

A two-by-two fixed in advance on PhantomWiki crossed an even share table (every article at $1/51$ of the wiki's relation facts) with the coverage line present or dropped. All four combinations are at most 1 of the 32 for every one of the 11 models under either target, so a true but irrelevant share table does not elicit the rule; whether a relevant one would is untested.

B.7 THE SOFTWARE-TESTING SETTING

Outside search, the bug form asks whether a bug of a stated type is present after a partial test run over modules whose share of past bugs is given, so the untested share plays the part of $1 - \delta$. The result, that the models on the $1 - \delta$ rule report the untested share as that probability, is stated in Sec. 3.1 and given per model in Table 10. The same replies replayed through the deployment pipeline are in Table 30 (App. H), and the trainings' transfer to this form is in Table 11 (App. F.5).

Table 10: **In the software-testing setting, every model on the $1 - \delta$ rule reports the untested share as the probability that a bug is present.** Share of parsed replies to the probability question, at the four partial test budgets, that equal the untested share of the program, per model, with the number parsed; the rule is the model's in the main benchmark.

Model	On the $1 - \delta$ rule	Untested share	Parsed
GPT-5.5	yes	100%	64 of 64
Grok 4.20 Reasoning	yes	100%	64 of 64
Kimi K2.7	yes	100%	62 of 64
Nova Pro	yes	100%	64 of 64
Sonnet 5	yes	100%	64 of 64
Haiku 4.5	yes	95%	64 of 64
Mistral Large 3	no	100%	63 of 64
Llama 4 Maverick	no	69%	64 of 64
Gemini 3.1 Pro	no	65%	60 of 64
GLM-5	no	45%	64 of 64
DeepSeek-V3.2	no	39%	64 of 64

Table 11: **Each training moves the same way on a task it never saw.** The $1 - \delta$ share of the untrained model and of each training, on NullSearch and on BugSearch at two seeds. NullSearch shares are pooled over its own 1,024 coverage-given replies. BugSearch shares are over the PARSED replies of 64 attempted per seed. Every share is over the replies that parsed, which is all 64 except the untrained Qwen, at 56 and 51. The two columns are different denominators and are not differences of one number.

	NullSearch $1 - \delta$		BugSearch $1 - \delta$	
Training	Untrained	Trained	Seed 3	Seed 4
Gemma, Brier	87%	0%	8%	5%
Gemma, log	86%	3%	41%	47%
Gemma, fixed-prior control	86%	0%	0%	0%
Qwen, supervised	78%	13%	22%	25%
Qwen, Brier	79%	98%	97%	94%
Qwen, Brier, coverage withheld	79%	88%	84%	88%
Gemma-2-9B-it, untrained	87%	—	92%	81%
Qwen2.5-14B-Instruct, untrained	78%	—	84%	84%

B.8 AMOUNTS AND COUNTS

Asked for an expected amount or count instead of a probability, after a partial check found nothing, models answer zero.

The *amount* is the expected total of incorrect invoices after a share δ of a stated total T was examined and no errors found; the *count* is how many faulty goods after a count that covered a share δ of a warehouse's bins and found none. The prompt fixes the unsearched part $(1 - \delta)T$; the field needs a

base rate π the prompt omits. Over 11 models, two instruments, four coverage levels and 32 episodes per combination, with and without a stated rate (5,632 responses), 2,601 of 2,611 answered no-rate responses are zero, the ten exceptions lie in two models, and no model reports the posterior at an assumed rate. The strong form of the default, the field filled with the unsearched part itself, transfers to amounts and counts in two of the 11 models, in eight responses at the two higher coverages; the weak form transfers in all: the field is filled with what the prompt fixes from the searched part, a zero.

C WHICH EVENT THE NUMBER DESCRIBES

This part tests the reading that $1 - \delta$ is one conditional probability reported in place of another, against the reading that it is a prompt habit with no event behind it. The detection claims of Sec. 3.2, the 256 of 265 event labels of Sec. 3.1 and the decline census of Table 13 rest on it. Unless stated, each condition has 16 episodes per model at provider default sampling, and conditions with 16 or 32 episodes are reported as counts without intervals.

C.1 THREE QUESTIONS IN SEPARATE CALLS

Asked in separate calls for the posterior and the miss likelihood, 18 of 30 models give the two the same number, and a second call comparing the two answers flags 19 of the 20 scored models on the $1 - \delta$ rule and none of the seven on the posterior rule.

Each episode asks three questions in separate stateless calls: the posterior $P(E \mid F)$, the miss likelihood $P(F \mid E)$, and the probability that the record is in the unsearched locations, which equals the posterior in this design and is reported but not scored. No coverage figure appears in the prompt, so equality across the questions cannot come from copying a stated figure. Pooled over the releases asked, 30 of 35 pass the calculation check, and 18 of those 30 return one number to the three questions, their median responses being equal (Fig. 6).

We also put the questions to 29 models with an assigned rule, 22 on the $1 - \delta$ rule and seven on the posterior rule, and score 27, two of the 22 being unscored (one accepts at most 4,096 output tokens where we request 6,000; every call to the other fails). The first two questions agree in 19 of the 20 scored on the $1 - \delta$ rule and differ in all seven posterior-rule models; the main text's criterion for agreement is two answers 0.02 apart or closer (Sec. 3.2).

Wider coverages (δ from 0.12 to 0.78), a second record type, fourteen open-weight releases and the three questions asked in one conversation with the posterior first all replicate the result; asking the likelihood first ends the equality, the models reporting δ in its place.

C.2 THE EVENTS NAMED BY SYMBOL

Defining the events in the prompt does not separate the two answers. With E (the record exists) and F (the searches find none) defined symbolically and $P(E \mid F)$ and $P(F \mid E)$ asked by symbol in separate calls, GPT-5.5 still returns 0.68 to both in most episodes, and Opus 4.8 keeps the two apart in every entry.

C.3 A STATED COVERAGE AGAINST THE EVIDENCE

Which coverage a model follows when a stated line disagrees with the evidence varies by model. On the enterprise corpus of Sec. 3.3, with the tabulated line shifted about 0.4 above or below the true coverage, GPT-5.5 follows an inflated line in every response and returns to the true value when the line is lowered (14 of 16), and an inflated line moves Opus 4.8 down as well (median shift -0.09).

C.4 BOTH EVENTS IN ONE REPLY

Asked for both events in one reply, most models on the $1 - \delta$ rule give them one number less often than in separate prompts, but leaving $1 - \delta$ is not reaching the posterior. On the 11 models quoted in the main text with rows ($\delta = 0.32$ and 0.78, 16 episodes each), eight of the ten this case covers give them one number more than 10 points less often than in separate prompts. Of the 96 responses

that leave $1-\delta$, 35 give the 0.5-prior posterior, 15 a reported zero, 45 something else, and 1 fails to parse.

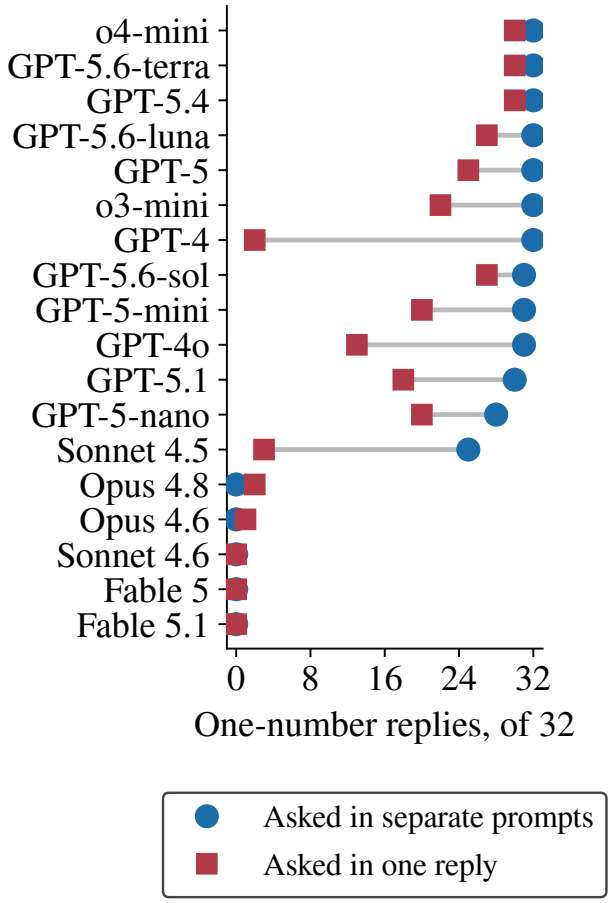


Figure 12: **Asked for the two events in one reply, all thirteen models on the $1-\delta$ rule give them one number less often, ten by more than 10 points.** Each row counts one model's replies, of 32, giving one number to existence and miss.

On the 19 wider-set releases that have rows, each measured against its own two-prompt rate (Table 12), a release counts as freed when its share of replies giving one number to both questions falls more than 10 points below that baseline. On that measure 10 of the 13 scored releases on the $1-\delta$ rule are freed, and the five posterior-rule releases give one number in none of 32 at baseline, so they cannot count as freed. Freed measures whether the two answers differ, not whether the existence answer leaves $1-\delta$.

C.5 The event named in the reply

Asked which event its number describes, a model on the $1-\delta$ rule mostly names the unsearched share or the miss likelihood, not the existence probability. In a second turn on the 11 models quoted in the main text with rows ($\delta = 0.32$ and 0.78, 16 episodes each), asked which event the probability just given referred to, seven of the ten models this case scores say the miss event or the unsearched share and three say the existence probability, among their responses at $1-\delta$; one model goes unlabelled under the strict parse.

A one-turn version puts the label in the same response as the probability, on the same 11 models with four schemas (32 coverage-given episodes each). With a required event field, the $1-\delta$ responses in the ten scored combinations are labelled the unsearched share in 184, the miss likelihood in 72

Table 12: **Asking both events in one reply makes the two answers differ in 10 of the 13 scored releases on the $1 - \delta$ rule.** Replies of 32 giving one number to both questions, in two prompts (baseline) and in one reply with a field for each. Freed means the count falls more than 10 points. The split is issued with the code. Not listed, scoring at neither form: o1. Posterior-rule releases give one number in none of 32 at baseline, so cannot be freed.

	one number, of 32			
Release	Baseline	One reply	Fall	Outcome
GPT-4	32	2	-94	freed
Sonnet 4.5	25	3	-69	freed
GPT-4o	31	13	-56	freed
GPT-5.1	30	18	-38	freed
GPT-5-mini	31	20	-34	freed
o3-mini	32	22	-31	freed
GPT-5-nano	28	20	-25	freed
GPT-5	32	25	-22	freed
GPT-5.6-luna	32	27	-16	freed
GPT-5.6-sol	31	27	-12	freed
GPT-5.4	32	30	-6	not freed
GPT-5.6-terra	32	30	-6	not freed
o4-mini	32	30	-6	not freed
Fable 5	0	0	$+0$	0 at baseline, cannot fire
Fable 5.1	0	0	$+0$	0 at baseline, cannot fire
Sonnet 4.6	0	0	$+0$	0 at baseline, cannot fire
Opus 4.6	0	1	$+3$	0 at baseline, cannot fire
Opus 4.8	0	2	$+6$	0 at baseline, cannot fire

and the existence probability in 9 of 265, the 256 of 265 of Sec. 3.1. The field moves Gemini 3.1 Pro, a posterior model under the plain question, onto $1 - \delta$ in every case, and six of 11 models are internally consistent in every complete (probability, label, prior) row.

C.6 SELF-CHECKING AND CROSS-JUDGING

Told to re-examine the probability it has just given against its own $P(F \mid E)$ and Bayes' rule and then reply again, GPT-5.5 reaches the 0.5-prior posterior in 42 of its 48 replies over three runs of 16, and GPT-5.6-sol in all 16 of its replies. Given another auditor's reply, GPT-5.5 keeps $1 - \delta$ until the Bayes prompt is added.

C.7 THE DECLINE OPTION CENSUS

Offered a structured decline (output `{"determinable": false}` if the probability cannot be determined) on the main partial-search question, 13 of the 40 models with this instrument decline on at least 60% of partial searches with coverage stated, and withholding the coverage moves the rates in both directions (Table 13; 64 responses per model in each coverage combination, never pooled). Sonnet 5 and Kimi K2.7 return nothing parseable at the survey's 2,000-token budget and were given 8,000 tokens in their own cases: Kimi K2.7 then declines on all 64 partial searches and Sonnet 5 on 9 of its 63 parsed.

D OUTPUT FORM, BASE RATE AND PRIOR

This part shows what the reply form and the prompt's inputs do to the number a model reports after a failed search. Given the base rate, or asked for the prior it assumes, a model on the $1 - \delta$ rule reports the posterior instead; an event field beside the probability field leaves that correction in place while the probability field still asks whether the record exists, and rewording that field brings the substitution back. The main text's claims that the substitution is a missing input and not a missing ability (Sec. 3.1), that most models name the event of their $1 - \delta$ report (Sec. 3.1), and that a field listing the missing inputs leaves the substitution in place (Table 18) rest on it, and the reference prior of 0.5 (Sec. 2) on its last subsection.

D.1 THE BASE RATE SUPPLIED IN THE PROMPT

Once the base rate is in the prompt, models on the $1-\delta$ rule use it, whether it arrives as an instruction or as a fact.

With an explicit prior line at $\pi \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, at coverage levels 2 and 4, 16 episodes per combination, on the nine models with proxy endpoints, a response adopts the prior when it is within 0.02 of $p^*(\delta, \pi)$, pooled over nine of the ten combinations; the tenth admits a small response and the correct response equally and is out of the counts. Four of the nine adopt in 144 of 144 responses, one in 131 of 144 and one in 34 of 142. The three models on neither rule adopt in fewer than a third of theirs (19, 31 and 12 of their 144, 143 and 144).

A base rate inserted as a fact with no instruction to use it ("Company-wide, 1,500 of 5,000 employees have a parking_permit record") is adopted as well. Both tested models give the posterior at the stated company-wide rate, $p^*(\delta, 0.3)$, in 16 of 16 episodes in each condition, GPT-5.5's responses at $\delta = 0.32$ being 0.2256637168, the posterior to ten decimal places, so what was missing was an input and not calculation ability. Of 34 evaluated models, 20 adopt the stated-rate posterior in at least 60% of responses. One failure mode is to copy the rate itself: GPT-5.1 answers 0.3 to the digit in 11 of 32 responses spread across all four coverages, a copy and not uptake because the value is unaffected by coverage.

The same holds away from search. A screening record puts the detection likelihood in the prompt and keeps the prevalence out of the transcript (2,112 calls over three reply forms, two of them read here), and each of the 11 models is elicited 64 times with and without one inserted sentence stating the prevalence. Without it, each stays beneath 60% on the miss likelihood, and the probability instead takes the posterior at a 0.5 prior, or 0, or something the rules do not predict. With the

Table 13: **Decline survey.** Every model given the decline-permitted question at the four partial-coverage levels with coverage disclosed (last column: coverage withheld, never pooled). A decline is the structured output itself, and unparsed replies never count as declines. A model declines at a rate of at least 60%: 13 of 40 models overall, and 6 of 13 among the releases fixed before the counts. Last rows: the two models reachable only at 8,000 tokens, never pooled; with them, 14 of 42 overall and 7 of 15 models decline.

Provider	Release	Declines	Numeric	Rate	Rate, withheld
Alibaba	Qwen3-8B	46	18	72%	92%
	Qwen3-14B	41	23	64%	100%
	Qwen3-32B-AWQ	22	42	34%	78%
	Qwen3-32B	19	45	30%	72%
	Qwen3-32B-FP8	17	47	27%	69%
	Qwen2.5-72B-Instruct	14	50	22%	39%
	Qwen3-30B-A3B	0	64	0%	3%
Amazon	Nova Pro	19	45	30%	31%
Anthropic	Opus 4.8	64	0	100%	41%
	Fable 5.1	62	2	97%	100%
	Fable 5	56	8	88%	81%
	Haiku 4.5	11	53	17%	16%
	Sonnet 4.5	8	56	12%	8%
DeepSeek	DeepSeek-V3.2	16	48	25%	38%
Google	Gemini 3.1 Pro	64	0	100%	100%
	Gemma-3-12B-it	40	24	62%	41%
	Gemma-2-27B-it	2	62	3%	2%
Meta	Llama 4 Maverick	19	45	30%	53%
Mistral	Mistral Large 3	10	54	16%	3%
OpenAI	GPT-5.5	64	0	100%	100%
	GPT-5.6-sol	64	0	100%	100%
	o1	56	8	88%	89%
	GPT-5-mini	43	21	67%	77%
	GPT-5-nano	39	25	61%	91%
	GPT-4 Turbo	29	35	45%	36%
	o3-mini	28	36	44%	47%
	GPT-4o-mini	24	40	38%	0%
	o4-mini	18	46	28%	53%
	GPT-4.1-nano	12	51	19%	9%
	GPT-4o	8	56	12%	5%
	GPT-5.2	8	56	12%	3%
	o3	3	61	5%	23%
	GPT-3.5 Turbo	1	63	2%	8%
	GPT-4	1	63	2%	0%
	GPT-5.4-mini	1	63	2%	2%
	GPT-4.1	0	64	0%	0%
	GPT-4.1-mini	0	64	0%	0%
	GPT-5.4-nano	0	64	0%	0%
xAI	Grok 4.20 Reasoning	64	0	100%	100%
Zhipu	GLM-5	28	36	44%	41%
Moonshot	Kimi K2.7 (8,000 tokens)	64	0	100%	100%
Anthropic	Sonnet 5 (8,000 tokens)	9	54	14%	18%

Table 14: **Given the base rate, 12 of the 18 models on the $1-\delta$ rule adopt it.** Models that adopt the base rate the prompt gives, in at least 60% of responses, by the rule they follow when no base rate is given.

Rule without a prior	Adopt the ambient rate
$1-\delta$	12 of 18
mixed	3 of 11
posterior	5 of 5

prevalence in the transcript, seven models report the posterior at that prevalence in 64 responses, and nine of the 11 convert it. Declines are absent across the 2,112 calls.

D.2 THE PRIOR THE MODEL ASSUMES

A required field for the prior the model assumes converts half of the ten scored models to the posterior at the prior they state, and the models that keep $1 - \delta$ fill the field with 1.0.

The one-turn cases run on the 11 models the main text quotes, with 32 coverage-given episodes each, and a model is scored in a case only where its unparsed replies stay within the quarter set in advance. With a required assumed-prior field, six of the 11 models convert to the posterior at their own prior, or five of the ten the main text scores: four in every one of their 32 cases, one in 30 of 32 and one in 28 of 32, while the two that keep $1 - \delta$ fill the prior field with 1.0 in 14 and 28 of 32 cases. With the field renamed to the probability of a miss given existence, nine of the 11 models give $1 - \delta$, the correct one, in every response, and the other two in most of their responses (24 and 23 of the 32).

With the event label and the prior required together, on the same episodes, three of the converting models convert far under their count with only the prior, whereas two convert with the label present, and the five near zero under only the prior stay there (Fig. 8). On fourteen models of the wider set, asking for the prior converts nine of the thirteen scored (Table 15); requiring the event as well lowers the conversion in eight of the nine, raises one from 28 to 30, and takes only GPT-5.6-sol down to 1.

Table 15: **The prior field, and the same field with the event field added.** Replies of 32 giving the posterior at the prior the reply states, over the wider set's fourteen. 9 of the 13 scored reach 24 with the prior field. Adding the event field, which rewords the probability field, lowers that in 8 of the 9, raises it in 1, and takes only GPT-5.6-sol under 8. The prior-field column is this run's own count; the Record column of Table 16 is the earlier count of record, and the two differ by at most 3 per release. o1 is not scored at either: 15 and 25 of its replies do not parse, past the quarter set in advance.

Release	Prior field, this run	With event field	
GPT-4	0	0	same, never converted
GPT-4o	3	0	lower by 3, never converted
GPT-5	31	12	lower by 19
GPT-5-mini	32	27	lower by 5
GPT-5-nano	28	30	rises
GPT-5.1	2	0	lower by 2, never converted
GPT-5.4	0	0	same, never converted
GPT-5.6-luna	30	20	lower by 10
GPT-5.6-sol	32	1	lower by 31, fell under 8
GPT-5.6-terra	30	22	lower by 8
o3-mini	32	20	lower by 12
o4-mini	30	21	lower by 9
Sonnet 4.5	31	22	lower by 9

D.3 THE EVENT FIELD BESIDE THE PRIOR FIELD, AND THE REWORDED PROBABILITY FIELD

What returns the substitution is the rewording of the probability field, not the event label next to it.

The five reply forms went to the 21 rule-classified models outside the 11 scored above, reported as their own group and not pooled with the 11. Where the reply asks for the probability and nothing else, the assignment is unchanged: each of the 14 on the miss likelihood reports it in at least 25 of its scored replies and each of the five classed to the posterior reports it in none.

The matched form separates the two things a reply form with an event field changes at once, adding the label and rewording the probability field. When the label field is added but the probability field

still asks whether the record exists, the correction survives in eight of the nine wider-set models the prior field corrects on its own, at 26 to 32 of 32 (Table 16); Sonnet 4.5 is the ninth, at 21, below that line and well above a loss. So the weakening under the two-field form belongs to the form that redefines the number as "the number you report" rather than to the second field as such. Eight of the nine are one provider's, so this is bounded in the same way the rest of the wider set is. Over all fourteen models, the five scored and the nine of the wider set, the totals are: 418 of 448 responses give the posterior under the matched form (156 of 160 and 262 of 288), and 266 once the probability field is reworded (91 of 160 and 175 of 288). For the nine, the two counts come from separate runs, so the fall from 418 to 266 compares two collections rather than three conditions of one run.

Table 16: **The matched reply form on the nine wider-set releases the prior field corrects on its own.** Replies of 32 giving the posterior at the prior the reply states, under a label field next to a probability field that still asks whether the record exists. Rerun is the prior field by itself, record is its earlier value, and drift is the gap between them, at most 2. 8 of the nine reach 24 of 32 and Sonnet 4.5 is between, at 21. 8 of the nine are one provider's.

Release	Matched	Rerun	Record	Drift
GPT-5-mini	32	32	32	0
GPT-5-nano	32	30	30	0
GPT-5.6-terra	32	32	32	0
GPT-5	31	31	30	1
GPT-5.6-luna	30	31	32	1
o4-mini	30	29	30	1
o3-mini	28	31	29	2
GPT-5.6-sol	26	31	32	1
Sonnet 4.5 (between)	21	30	28	2

D.4 What each reply form asked for

Six one-turn reply forms were put to a failed search (Table 17), each reported in Sec. 3.1. The definitions are the prompt templates' words.

Table 17: **What each reply form asked for.** The probability field's own definition, whether the model chose the event, and where the prior came from.

reply form	what the probability field asked for	the event	the prior
strict existence (`prob`)	the probability that the record exists anywhere in the company systems	set by the wording	none asked
event field (`prob_event`)	the number you report, with a companion field naming which quantity it is	the model's choice of three	none asked
prior field (`prob_prior`)	the probability that the record exists anywhere in the company systems	set by the wording	the model's own
both fields (`prob_event_prior`)	the number you report	the model's choice of three	the model's own
matched form (`prob_exists_event_prior`)	the probability that the record exists anywhere in the company systems	the model's choice of three	the model's own
prefilled prior (`prob_prior_prefilled`)	the probability that the record exists anywhere in the company systems	set by the wording	supplied in the skeleton at 0.5

D.5 THE INPUTS-MISSING FIELD AND THE CLARIFICATION STEP

Naming the missing input is not the same as using it: the models on the $1 - \delta$ rule that name the prior, in an inputs-missing field or in a clarification step, go on reporting $1 - \delta$ in most of the same replies.

An inputs-missing field accompanied the probability in each order (2,112 calls). Four releases name the prior in at least 60% of their responses that carry the field, and the two of them on the $1 - \delta$ rule, GPT-5.5 and Sonnet 5, go on reporting it in most of the same responses. The field moves the probability in both directions: GPT-5.5 falls from every baseline response on $1 - \delta$ to 65% with the field first, and Llama 4 Maverick rises from 14% to 41%. The keyword rule that reads the field fails in one direction only, so every naming share is a lower limit.

A clarification step leaves the substitution in place as well. Each model first lists the inputs the requested probability needs and the prompt withholds, proposes a value for each, and then answers under each proposal in turn and under none, over sixteen episodes at each coverage level. Where the prompt withholds the base rate, the two columns of Table 18 do not move together: GPT-5.5 and Gemini 3.1 Pro both name the base rate in 64 of 64 clarification replies, and GPT-5.5 then gives the miss likelihood in every scored reply while Gemini gives it in five of 64 and the posterior at a 0.5 prior in the other 59, and Nova Pro and Haiku 4.5 name it in none and give the miss likelihood in almost all. The value a model proposes is used once it is stated: where the proposed input is the base rate, the probability that follows is the posterior at that proposed value in 308 of 321 replies and the miss likelihood in two.

Table 18: **Naming the missing input and then filling the field are separate.** Per release, over the episodes with the base rate withheld: how often the clarification step names the base rate among the inputs it lists, and how often the probability field then holds the unread share. Denominators differ because empty and unreadable replies enter neither. Replies cut off at the store's limit are not excluded.

Release	Names the base rate	Reports the unread share
Kimi K2.7	58 / 58	38 / 38
GPT-5.5	64 / 64	64 / 64
Gemini 3.1 Pro	64 / 64	5 / 64
Sonnet 5	63 / 64	44 / 64
Grok 4.20 Reasoning	34 / 64	64 / 64
Llama 4 Maverick	24 / 64	5 / 64
Mistral Large 3	13 / 63	3 / 63
DeepSeek-V3.2	7 / 64	31 / 60
GLM-5	1 / 64	18 / 62
Nova Pro	0 / 64	60 / 61
Haiku 4.5	0 / 64	62 / 64

D.6 IMPLICIT-PRIOR ANALYSIS

No fixed implicit prior explains the $1 - \delta$ response. The three-quantity test (App. C.1) shows within-episode that the substituted value is the likelihood, not a prior-bearing posterior, and the measurements here test the remaining implicit-prior explanations.

Each episode shows the model a search log and asks for the probability that the record exists. A second turn asks what prior it assumed beforehand, in two question forms, the first permitting a decline and the second urging one. Opus 4.8 names 0.5 in all 80 episodes, with 79 of the 80 equal to the posterior at the assumed prior. The $1 - \delta$ models, which report the miss likelihood rather than the posterior, cannot name a consistent prior under either wording: permitted a decline, 73 of 96 replies say they have no base rate and 23 give 1.0, at which the posterior is 1 at every δ, and urged for one, GPT-5.5 and GPT-5.4 give in 59 of 64 episodes a prior that contradicts what they

themselves reported. The unsearched and zero-coverage conditions also change the answer, to 0.5 with no search log and to near one once a zero-coverage search is in the prompt (App. E.2), and a prior kept before the search cannot change with search effort. The prior is an input these models lack, and once it is in the prompt they use it (Sec. 3.1).

E Robustness and alternative explanations

This part collects the tests that rule out other readings of the $1 - \delta$ response, and the calibration census. The account of the default as a widespread rule (Sec. 3.1) and the statement that no calibration score can say which event a report describes (Sec. 3.3) rest on it. Two rivals are tested elsewhere: no fixed implicit prior fits the responses, under either the prior grid or the direct elicitation (App. D.6), and coverage estimation is a different step from the conversion, because stated and summed coverage can be compared directly (App. F).

E.1 A stated number copied into its complement

Both rules survive changes to how the shares are written and to which share table they describe. With the share percentages written as words, 44 of GPT-5.5's 64 responses equal one minus the summed shares to six or more decimal places and the rest lie within 0.02 of $1 - \delta$; with the shares stated to six decimal places, it answers 0.907349, the complement of a sum that appears nowhere in the prompt. On four kept-out share tables for new record types, GPT-5.5 gives $1 - \delta$ in 127 or 128 of 128 responses, and Opus 4.8 the 0.5-prior posterior in 127 of 128.

E.2 An artefact of the wording or the layout

A generic reminder that absence of evidence is not evidence of absence, a permission sentence stating that a mid-range answer is acceptable, either of its two ablations, and instructed in-format reasoning leave the substitution in place.

Varying one feature of the elicitation at a time leaves each model's rule where it was, with one exception. Under an imperfect search with a stated hit rate ρ, GPT-5.5 reports $1 - \rho\delta$ and Opus 4.8 the posterior built with ρ. Dropping the opening clause or substituting the auditor persona moves no model's $1 - \delta$ share by more than 20 points, and GPT-5.5 returns the same response at low and high reasoning effort in every twinned episode. The exception is the frequency instrument: with the frequency format, the model that computes the posterior reports the miss likelihood instead, the opposite of the human literature, where frequency formats reduce the transposition (McDowell & Jacobs, 2017). The counts for each condition are in App. I.

Further perturbations move at most the zero-coverage answer. Twelve rephrasings of the question, entity names, seeds, new takes, and reruns of the identical configuration leave the rule undisturbed in all twinned responses, with the zero-coverage answer changing for API-hosted models and not for open-weight ones. The response persists with a plain number and no JSON, and under a smaller forced-choice elicitation each model picks its own rule's nearest option in nearly all responses.

E.3 Confidence on the wrong reports

Table 19 gives each model's confidence on its wrong reports and on its right ones. Pooled, a confidence ranks reports against the event it names at an area under the ROC curve of 0.948 with the base rate given and 0.879 without it, and against the other event at 0.365 and 0.334, both below chance.

E.4 Calibration cannot identify the event

Each model is calibrated exactly where the deployment's base rate agrees with its rule, since $p^*(\delta, \frac{1}{1+\delta}) = 1 - \delta$ and the 0.5-prior responses have almost no calibration error at $\pi = 0.5$, and otherwise each is miscalibrated. In the deployment its rule implies, a model's responses are indistinguishable from the posterior, so a calibration check run there cannot tell which event they refer to.

Table 19: **Confidence on the wrong report against the right one, release by release.** At the planted base rate. A report is wrong for the asked question when it is more than 0.02 from the posterior. Repeats: confidence replies repeating the reported number, excluded. Dashes: no report of that kind. Pooled, a confidence ranks reports against the event it names at an area under the ROC curve of 0.948 with the base rate given and 0.879 without it, and against the other event at 0.365 and 0.334, both below chance.

	reports		mean P(True)			
Release	wrong	right	wrong	right	diff.	repeats
DeepSeek-V3.2	11	0	0.547	–	–	24
GLM-5‡	26	1	0.309	1.000	−0.691	9
GPT-5.5	16	2	0.000	0.500	−0.500	14
Gemini 3.1 Pro	1	2	1.000	1.000	+0.000	29
Grok 4.20 Reasoning	9	11	0.667	1.000	−0.333	12
Haiku 4.5	1	1	0.680	0.226	+0.454	39
Llama 4 Maverick	4	3	0.429	0.226	+0.203	26
Mistral Large 3	12	0	0.362	–	–	22
Nova Pro	0	0	–	–	–	48
Sonnet 5	10	15	0.780	0.993	−0.213	10

Matched over 7 of the 10 releases. Repeats is out of 64. ‡: the substituted rows of this release are reported, not classified.

Table 20 repeats the computation for the 29 models assigned a rule, 22 on $1-\delta$ and 7 on the posterior, under four base rates. The expected calibration error takes ten equal bins on the reported probability and, in each bin, the absolute difference between the mean report and the mean true posterior $p^*(\delta_i, \pi)$ of its rows, weighted by the bin's count. The $1-\delta$ group averages 0.36 at a base rate of 0.3 and 0.26 at 0.9, the posterior group 0.15 and 0.47; at $\pi = 0.5$ the two average 0.20 and 0.02, and in the last column, where the base rate equals $1/(1+\delta)$ at each coverage, the order reverses to 0.01 against 0.23, so a low value there is rule adherence restated, not calibration skill.

F Mechanism: Coverage by Counting, Scale, Training, and Derivation Traces

This part asks where in a model's computation the substitution happens and what moves it. The coverage is computed correctly and the error is at the step that turns it into a probability; six of the 11 models build a coverage by counting locations when none is given and apply the same rule to it; inside two open models a linear probe reads the stated prior and an activation splice moves one of them to the posterior; the substitution strengthens with scale and is in place in base models before post-training; and a reward for the posterior changes how often the three trained models report $1-\delta$ but installs the posterior in none of them on an unseen wording. The main text rests on it for the counting result (Fig. 4), the probe and splice, the scale result (Fig. 5), the 13 base and instruct pairs, and the training result (App. F.5).

F.1 Where the substitution starts

The step that decides the rule is the conversion, not the coverage estimate. Asked directly for its covered share, a model works it out correctly and then converts that share into $1-\delta$ and not into a posterior, whether the coverage total is given or the model sums the shares itself (Table 6).

Forced derivations show the same thing in the model's own words. The forced-derivation prompt requires the model to list the searched locations, say what the failed searches show, and derive its probability before its number, on 16 episodes per model at $\delta = 0$, $\delta \approx 0.33$ and $\delta = 1$. Models asked to derive their answer first mostly reason without mentioning a base rate and report $1-\delta$. Across 11

Table 20: **Expected calibration error of every assigned release under four base rates, computed from the committed replies.** Reliability is exact under the search model, so no outcome sampling is needed. The last column is the world whose base rate equals $1/(1+\delta)$ at each coverage: there a $1-\delta$ reply coincides with the correct posterior, so a low value in that column is rule adherence restated, not calibration skill. At $\pi = 0.5$ the 22 $1-\delta$ releases average 0.20 and the 7 posterior releases 0.02; at $\pi = 1/(1+\delta)$ the order reverses (0.01 against 0.23).

Release	Rule	n	$\pi = 0.3$	$\pi = 0.5$	$\pi = 0.9$	$\pi = 1/(1+\delta)$
GPT-5-nano	$1-\delta$ rule	64	0.30	0.15	0.32	0.07
Sonnet 4.5	$1-\delta$ rule	64	0.31	0.15	0.31	0.06
Kimi K2.7	$1-\delta$ rule	123	0.32	0.16	0.30	0.05
Sonnet 5	$1-\delta$ rule	288	0.32	0.16	0.30	0.07
Nova Pro	$1-\delta$ rule	64	0.35	0.21	0.28	0.03
GPT-4 Turbo	$1-\delta$ rule	64	0.36	0.21	0.27	0.02
Haiku 4.5	$1-\delta$ rule	128	0.36	0.20	0.26	0.01
GPT-5-mini	$1-\delta$ rule	64	0.36	0.21	0.26	0.01
o3-mini	$1-\delta$ rule	64	0.37	0.21	0.25	0.00
GPT-5	$1-\delta$ rule	121	0.37	0.21	0.25	0.00
GPT-5.1	$1-\delta$ rule	128	0.37	0.22	0.25	0.00
GPT-5.6-luna	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
GPT-5.6-terra	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
DeepSeek-R1	$1-\delta$ rule	128	0.37	0.22	0.25	0.00
GPT-5.6-sol	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
GPT-4o	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
GPT-4	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
o1	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
o4-mini	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
Grok 4.20 Reasoning	$1-\delta$ rule	64	0.37	0.22	0.25	0.00
GPT-5.5	$1-\delta$ rule	738	0.37	0.22	0.25	0.00
GPT-5.4	$1-\delta$ rule	288	0.38	0.22	0.24	0.00
Sonnet 4.6	posterior	64	0.13	0.05	0.51	0.26
Opus 4.6	posterior	64	0.14	0.01	0.48	0.23
Fable 5	posterior	64	0.15	0.00	0.46	0.22
Fable 5.1	posterior	64	0.15	0.00	0.46	0.22
Opus 4.8	posterior	551	0.16	0.00	0.46	0.22
Gemini 3.1 Pro	posterior	60	0.16	0.00	0.46	0.22
Llama 4 Maverick	posterior	64	0.17	0.05	0.47	0.22
Mean, $1-\delta$ rule (22)			0.36	0.20	0.26	0.01
SD, $1-\delta$ rule (22)			0.02	0.02	0.02	0.02
Mean, posterior (7)			0.15	0.02	0.47	0.23
SD, posterior (7)			0.01	0.02	0.02	0.02

models at 16 traces each, the five on $1-\delta$ state a prior in 1, 1, 2, 7 and 11 traces, traces with none report $1-\delta$ in 94% of cases and traces with one in 45%, and Gemini 3.1 Pro states one every time and reports the posterior at it. At informative coverage the rule persists, GPT-5.5 giving $1-\delta$ in 16 of 16 derived responses and Opus 4.8 the 0.5-prior posterior in 16 of 16; only the zero-coverage extremes soften toward 0.5 (GPT-5.5's certainty responses from 86% to 41%).

A representative GPT-5.5 derivation says correctly what the failed search shows and where it is silent, and still reports $1-\delta$.

> The exhaustive null scans establish that no parking_permit record for Omar Okafor is filed in the scanned candidate locations totaling 22% + 10% = 32% of historical filing share. They do not establish absence from the unscanned candidate locations: facilities/parking (40%), facilities/badge_events (16%), benefits_portal/enrollments (7%), and travel_desk/bookings (5%), totaling 68%. Using the filing-share coverage as the probability basis, the remaining probability that the record exists somewhere unscanned is 1 - 0.32 = 0.68.

Forcing the derivation fixes only the zero-coverage extremes and never the conversion, so the number a model reports is not governed by what it writes about the evidence: GPT-5.6-sol and Fable 5.1 both write down what the failed searches show in 16 of 16 derivations, and GPT-5.6-sol reports $1-\delta$ in all 16 while Fable 5.1 derives the 0.5-prior posterior in 16 of 16.

F.2 Coverage by counting

Given the prior and a list of locations but no coverage figure, six of the 11 models count the searched locations and report the posterior at that count, so the counting rule is not another kind of model but the same rule under a second prompt.

The mirror of the main case gives the prior and withholds the coverage: the prompt lists six candidate locations without shares, shows no coverage figure, and shows a scan log in which s of the k listed candidates appear. The 11 models answered 64 episodes each (704 responses, 2 unparsed). Because the counting coverage s/k and the true, share-weighted coverage are equal at one level and the counting posterior is within 0.04 of the given prior at the lowest rate, the case can tell the three accounts apart on 440 of the 704 responses (40 per model), and only those count. Solving each response for the miss likelihood that would make it correct returns $2/3$, $2/3$, $2/3$ and $1/2$ by level, which is $1-s/k$ for $s=2,2,2,3$ searched of the $k=6$ candidate locations. This recovery was made from the responses only, before any account was supplied, and the scoring below confirmed it.

A prospective test then varied the ratio s/k over three ratios at $k=6$ and $k=12$ (528 responses, 48 per model, 2 unparsed). Its three counting targets are at least 0.05 from the true-coverage posterior and from the prior at every rate. The six counters hold at 93.8 to 100% in the ratio test, are within 0.02 of the counting posterior in 276 of 288 responses, and in 138 of 144 at each list length. A line-count account (p^* at $(s+3)/k$) predicts 0 of their 288, the true coverage 0, and the given prior 4. Of the five that do not count, four almost never do (at most 2 of 48): one returns the given prior, two lean that way, and one fits no account. Haiku 4.5, which skips counting, is the only model the list length moves; its counting share rises 37.5 points from six to twelve locations.

Doubling the list at a fixed coverage changes nothing (138 of 144 responses at six and at twelve locations), so what the six build is the fraction k/N of the list searched, an estimate of the coverage, and the rule they then apply to it is the same rule as before. The error is in what they count and not in how they combine it with the base rate: counting treats the locations as equal in record share.

F.3 The probe and the activation splice

The splice of the episode's own stated prior was run on Qwen3-32B and moves it to the posterior; on Gemma-2-9B-it only the other-prior splice was run, and it leaves the model on $1-\delta$, though a probe reads the prior in both.

On Gemma-2-9B-it with the coverage line in the prompt, a linear probe reads the stated prior at fixed coverage at blocks 18 to 24 (three-class accuracy 1.000 in both levels on two independent scorers). Splicing the prior sentence's positions from a prompt with a different prior moves at most 2% of 128

responses off $1-\delta$ at any single block. The random-position splice moves none of 5,376 (one seed, two independent scorers).

On Qwen3-32B the probe reads the prior at fixed coverage at or above 0.95 over 19 consecutive blocks at one level and 11 at the other (two independent scorers). The same splice moves 79 of 128 responses to the posterior at the episode's own prior at block 17, 60% or more over blocks 15 to 19 and at blocks 0 to 20 and 22 at the lower coverage. The random-position and other-prior controls move none of 8,192 to the episode's posterior at any block, and none from block 46 on, where the probe reads the prior best (one seed, two independent scorers and a tally). A replication sends at least 60% of the replies at each block from block 0 to block 20 and decays after it, and the deepest coverage stays beneath that at each block in either (two seeds, three reads).

Three caveats bound the reading. A splice at block 0 is a splice at the input, so a prior placed in the prompt by another means is not excluded. The random-position splice produced a much smaller disturbance than the main splice, so its lack of movement says little of position; a matched-amplitude control settles that, since writing the same activations at as many random positions (amplitude ratio 1.12) moves at most 0.79% of replies to the target at any block, against 60% or above for the treatment from block 0 to block 20. And the other-prior source moves the response to its own posterior in at most 7 of its 128 at each block, 2 to 6 over blocks 15 to 19, where the main splice sends 79 to the episode's posterior at block 17; the effect is absent at a second prior in these episodes, and whether that is specificity to the prior or to the points it was taken at is undecidable from them.

F.4 Scale

The $1-\delta$ rule appears at the step that turns coverage into a probability at smaller model scales than coverage estimation, and, in the base-and-instruct pairs we measured, the mapping from a given coverage to $1-\delta$ is present before post-training.

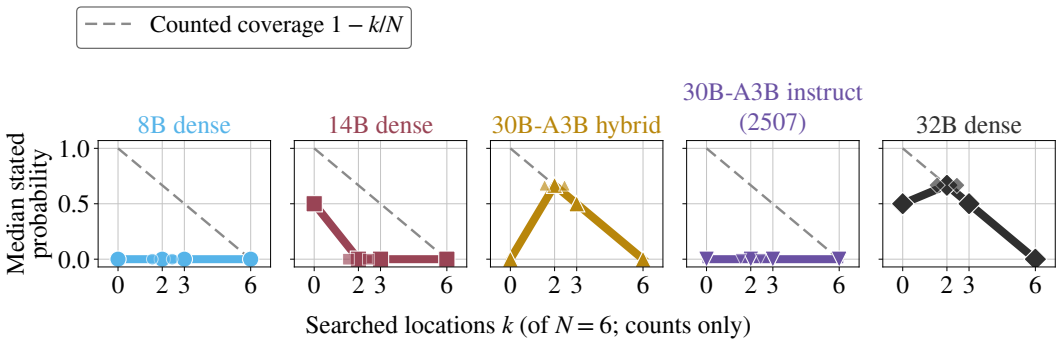


Figure 13: **Only two of the five releases follow the counting rule.** With shares withheld, the 32B dense and 30B-A3B hybrid track $1-k/N$ at $k=2$ and $k=3$; none meets it at $k=0$. Median stated probability against the searched-location count k, one part per release, against the counting rule (gray dashed), not the counting posterior at k/N. Both 30B-A3B releases are the same size; the instruct (2507) answers 0 throughout, as do 8B and 14B dense apart from 14B at $k=0$. Side markers at $k=2$ are the other two levels that also search two locations.

Converging to $1-\delta$ means converging to the substituted quantity, not gaining a capability. With the coverage supplied, the substitution strengthens with scale as the coverage summing beneath it does (Fig. 5). The plot holds 20 open-weight points, 1.0B–72B: 15 of the 19 instruct models in its data file, in the Qwen and the two Gemma families, and five of the six Qwen2.5 base ladder points, one excluded by the 25% unparsed-response allowance.

Base and instruct pairs. Thirteen complete base and instruct pairs from five families answered the main question with the coverage given at the four partial levels (16 episodes per level; the four newest-generation pairs are read at two levels). The instruct models are on $1-\delta$ at a rate that rises with size inside Qwen2.5, Gemma-3 and OLMo-2, from none under 4B (at most 9% of 32 responses), through 16 to 34% at 7B and 8B, to 29 of 32 scored responses for Qwen2.5-14B-Instruct. The base models convert a given coverage at a lower rate than the instruct half in every comparison

that can be scored, ten of ten across nine of the thirteen pairs, Qwen2.5-14B contributing two of the ten. Qwen2.5-14B base, the only base model whose plain case can be scored, is at 36 of 64 attempted responses (56%) against 29 of 32 (91%) for its instruct half, and in 37 of 64 (58%) with an output frame.

The frame itself tells the output contract from the belief. Without it, twelve of thirteen base models leave more than half of their responses unparsed; with it, 11 of the thirteen fall under the 25% allowance and none gets worse, yet every framed base case stays far below its instruct half on $1-\delta$. A base model can be made to fill the numeric field reliably without acquiring the substitution. In these pairs the mapping from a given coverage to $1-\delta$ comes before post-training; whether the coverage estimate does is not settled at this sample.

F.5 Training

A reward for the correct posterior moves the $1-\delta$ share in a direction that reward and model set together, and installs the posterior in none of the three trained models on a wording they never saw; the same setup installs a fixed-prior curve, so the setup can install a curve.

Setup. Each training is read against the untrained model in the same cases, on two sampling seeds and two prompt wordings, with the criterion applied per case. Supervised fine-tuning of Qwen2.5-14B-Instruct uses the 3,072 coverage-given training episodes of the partial-coverage reward. Three instruct models from three providers were trained with a squared-error (Brier) reward against the oracle posterior $p^*(\delta, \pi)$, the prior in every training prompt, on the same 3,072 episodes (a quarter of the corpus) for three epochs (2,304 updates, a third of the budget). A Brier comparison trains the primary model on the corpus's coverage-withheld prompts, which carry the prior but no shares table or coverage summary. A clipped log-score reward on Gemma-2-9B-it, fixed in advance, tells the outcome from the choice of proper score. The optimizer is GRPO (TRL 1.10.0) with LoRA (Hu et al., 2022) rank 32 and alpha twice the rank on the attention and MLP projections, learning rate 10^{-5}, a KL penalty of 0.04 to the starting model, and 4 prompts in each optimizer step, each with 8 generations. A positive control, the same reward against a fixed-prior target with no prior in the prompt, tests whether the setup can install any curve.

Criteria. The criteria are stated before the numbers: a trained model removes the default if its $1-\delta$ share in the no-prior case falls to at most half its untrained baseline on the matched case and its 0.5-prior posterior share exceeds its $1-\delta$ share; it installs the posterior if, given a prior, at least 60% of its responses lie within 0.02 of $p^*(\delta, \pi)$ on unseen environments and an unseen wording; one meeting the second without the first has learned the formula. One whose no-prior $1-\delta$ share rises above its baseline without a rise in its posterior share has reinforced the default, with any movement in conversion reported alongside. The cases that can decide conversion exclude the 0.05-prior level, where the posterior lies within 0.02 of the prior itself and conversion cannot be told from a repeat of the prior.

Capacity of the setup. On unseen environments Gemma-2-9B-it reaches the fixed-prior target in every response and OLMo-2-13B-Instruct in 98.8% (506 of 512), so their outcomes below are claims about the reward, whereas Qwen2.5-14B-Instruct reaches it in 8 to 13%, so every Qwen outcome below is bounded by the setup.

Supervised training. With supervised fine-tuning the $1-\delta$ share in the no-prior case goes from 87 to 13% on the training wording and from 67 to 14% on an unseen wording, with the 0.5-prior posterior share between 6 and 8% in every case. What replaces $1-\delta$ is a steeper curve, with medians on or above $1-\delta$ at the two shallower levels and far below it at the two deep levels, not a removal. Given a prior it adopts the posterior in 1.0% of responses on the cases that can decide it (10 of 1,024, against 16.8% untrained); the 0.05-prior cases would add 512 of 512 and are not in the count, so the supervised model learned that a low prior means a low probability, not the formula.

Brier training at partial coverage. Trained with a squared-error (Brier) reward against $p^*(\delta, \pi)$, the posterior at the episode's coverage and at the prior its prompt states, the three go three ways on $1-\delta$: removed in the first, reduced in the second, hardened in the third at LoRA rank 32 (Table 21). No trained model installs the posterior. Qwen2.5-14B-Instruct reinforces the default at rank 32,

reporting $1-\delta$ in 86 to 98% of no-prior partial searches across its two seeds in that evaluation against 64 to 90% untrained, with a posterior share of zero, while its conversion to the posterior at a given prior rises from 28 to 35% to 43 to 46% on the training wording and by nothing on the unseen wording. Gemma-2-9B-it removes the default completely, zero of 256 on $1-\delta$ in every case from 84 to 90% untrained, and installs not the posterior but a curve above the rule at every level, with conversion to the posterior at a given prior at 14.6% on the cases that can decide it, 25.4% on one phrasing and 3.7% on the other against 10% untrained on both. The comparison without the coverage summary on Qwen2.5-14B leaves the $1-\delta$ share where it started on the training template, raises the unseen template's share by about 20 points to the same level, and reaches the highest conversion to the posterior at the given prior of any trained model on the cases that can decide it, 26.8% (45.9% on one phrasing, 7.6% on the other). OLMo-2-13B-Instruct, whose untrained default is weaker (57 to 72% on $1-\delta$), removes the default on the training template at its final checkpoint at both sampling seeds ($1-\delta$ 14%, the 0.5-prior posterior 27 to 29%) and only partly on the unseen template (33 to 39% against 72%); its conversion to the posterior at a given prior on the cases that can decide it is 12.4% against 6.8% untrained. The clipped log-score reward on Gemma-2-9B-it reproduces the Brier result on the same data, budget and seed (1 to 4% on $1-\delta$, posterior zero), so the overshoot above the rule comes from the model or the data, not from the Brier score.

Conversion to the posterior at a given prior rises in every Brier-trained model on the cases that can decide it, by 3 to 10 points, far more on one phrasing than on the other; training never showed the models either phrasing.

Table 21: **A reward against the correct posterior, $p^*(\delta, \pi)$ at the episode's coverage and the prompt's prior, moves the substitution in a direction reward and model set together, installing the posterior in none of the three.** Percent of responses on the phrasing every training prompt carried. None reaches the criterion, the given-prior posterior share, on the unseen wording, where it is judged (Table 22). The labelled fit leaves its posterior under 8%. Qwen's baseline is its setup on the unseen wording, which bounds its range. Qwen's hardening is a rank-32 outcome and reverses at rank 128.

Trained model	reward	$1-\delta$	converts at the given prior	seeds
Qwen2.5-14B-Instruct	labelled	87 to 13	—	1
Gemma-2-9B-it	Brier	removed	83–90, 92.6–99.6	2
OLMo-2-13B-Instruct	Brier	reduced	below the target	3
Qwen2.5-14B-Instruct	Brier	hardened	22–52	3
fixed-prior baseline	—	—	8–13	—

The coverage-given phrasing. The 99.6% in Table 21 is a different phrasing, with the training prompts' coverage wording added. Re-phrased with the coverage line every training prompt carried (the cases that can decide conversion, with the 0.05-prior level out), Gemma-2-9B-it converts to the posterior at the given prior in 83.2 to 89.8% of responses across its four combinations (887 of 1,024 pooled) against 3.5% untrained on that phrasing, and its second training seed in 92.6 to 99.6% (996 of 1,024) against 3.3% untrained. Qwen2.5-14B-Instruct converts in 46.5 to 52.3% against 28.9% untrained, its second training seed in 21.9 to 28.1% and its third in 25.0 to 37.1%, a three-seed range of 24.1 points, each figure over 1,024 responses. OLMo-2-13B-Instruct converts in 0.4 to 8.6% (43 of 1,024) against 3.5% untrained, 89 of 1,024 at its second seed and 159 of 1,024 at its third, a three-seed range of 11.3 points over 1,024 responses, wider than the 11 points the confidence rule set before any seed ran, so this result's confidence is medium and not high. The 14.6, 26.8 and 12.4% figures above are the same models read on phrasings without that line. Table 22 gives the conversion per prior and range on each of the two phrasings, where no trained model reaches the criterion on the two priors that can decide it.

Rank 128. Retrained at LoRA rank 128, Gemma-2-9B-it repeats its rank-32 outcomes on the two criteria: the default is removed in the four combinations and the posterior is installed in not one of them. On Qwen2.5-14B-Instruct the posterior is still not in place at rank 128 (45 of the 1,024 responses at the two discriminating priors), but the two removal clauses now pass: its no-prior $1-\delta$

Table 22: **Trained models adopt a stated prior at a rate set by the prior and the prompt template.** The first three numeric columns give, at each stated prior π (the column head), the share of parsed responses within 0.05 of $p^*(\delta, \pi)$, and the last three the mean absolute error to it; each row pools the sweep's seeds, and the untrained row is the same pass's baseline. The template column names the prompt wording: training is the wording every training prompt used, unseen a held-out wording that training never showed, and the 60% criterion that no trained model reaches is judged on the unseen wording only.

model		template	0.05	0.30	0.70	MAE 0.05	MAE 0.30	MAE 0.70
Qwen2.5-14B Brier	untrained	training	77%	43%	24%	0.084	0.130	0.169
Qwen2.5-14B Brier	untrained	unseen	70%	34%	23%	0.158	0.181	0.199
Qwen2.5-14B Brier	trained	training	87%	73%	41%	0.066	0.039	0.075
Qwen2.5-14B Brier	trained	unseen	83%	36%	13%	0.096	0.142	0.238
Gemma-2-9B-it Brier	untrained	training	77%	39%	18%	0.162	0.133	0.182
Gemma-2-9B-it Brier	untrained	unseen	68%	46%	7%	0.118	0.100	0.251
Gemma-2-9B-it Brier	trained	training	100%	54%	35%	0.008	0.049	0.104
Gemma-2-9B-it Brier	trained	unseen	100%	49%	6%	0.010	0.051	0.107
OLMo-2-13B Brier	untrained	training	41%	35%	8%	0.187	0.112	0.298
OLMo-2-13B Brier	untrained	unseen	56%	19%	5%	0.215	0.191	0.307
OLMo-2-13B Brier	trained	training	69%	62%	22%	0.046	0.044	0.131
OLMo-2-13B Brier	trained	unseen	99%	28%	20%	0.020	0.087	0.218
Gemma-2-9B-it log	untrained	training	76%	40%	17%	0.174	0.129	0.182
Gemma-2-9B-it log	untrained	unseen	68%	46%	7%	0.116	0.093	0.253
Gemma-2-9B-it log	trained	training	100%	53%	45%	0.013	0.055	0.112
Gemma-2-9B-it log	trained	unseen	100%	53%	16%	0.012	0.051	0.112
Qwen2.5-14B SFT	untrained	training	77%	49%	26%	0.096	0.129	0.174
Qwen2.5-14B SFT	untrained	unseen	70%	37%	18%	0.147	0.156	0.218
Qwen2.5-14B SFT	trained	training	100%	59%	5%	0.006	0.046	0.093
Qwen2.5-14B SFT	trained	unseen	100%	53%	4%	0.008	0.050	0.095
Qwen2.5-14B coverage-withheld	untrained	training	77%	43%	23%	0.105	0.123	0.171
Qwen2.5-14B coverage-withheld	untrained	unseen	67%	40%	20%	0.177	0.161	0.200
Qwen2.5-14B coverage-withheld	trained	training	99%	98%	23%	0.012	0.014	0.176
Qwen2.5-14B coverage-withheld	trained	unseen	100%	41%	7%	0.010	0.066	0.220

and posterior shares are 0 to 0.04 and 0.12 to 0.26, against 0.97 to 1.00 and 0 at rank 32, a `seed 0` result against the three seeds of the rank-32 hardening.

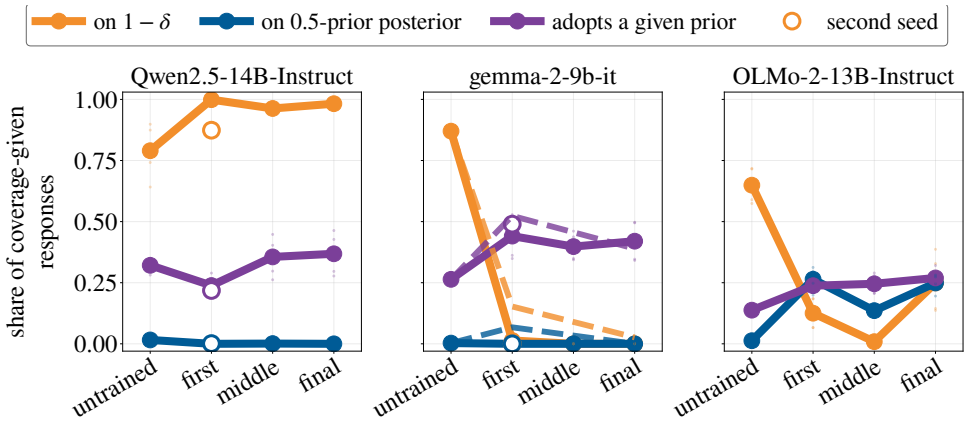


Figure 14: **Three Brier trainings, three paths: Qwen hardens the default, Gemma leaves it at the first checkpoint, OLMo leaves it and returns.** Share of coverage-given no-prior partial responses on $1-\delta$ (orange), on the 0.5-prior posterior (blue) and converting to the posterior at a given prior (purple), the share inside 0.02 of p^* when the prompt states the prior, at the untrained, first (ck460), middle (ck1380) and final checkpoints, pooled over four cases. Hollow markers are the second seed's first checkpoint and the dashed line is the log-score reward on Gemma.

Trajectories. First, middle and final checkpoints are observed for the three Brier trainings and first and final for the rest (Fig. 14). OLMo's path is not monotone: at the first and middle checkpoints the $1-\delta$ share is 12.5% and 0.9% pooled, and only in the last stretch of training does $1-\delta$ climb back to 36% on the unseen template. Gemma's removal is already complete at the first checkpoint (1.4% on $1-\delta$) and stays complete through the final (0.0%). The Qwen Brier model's $1-\delta$ share saturates by the first checkpoint (99.9% pooled). Three of the four trainings peak in conversion at the given prior at an early or middle checkpoint and decline, the closest approach, on the phrasing without the coverage line, to the 60% criterion on record being the log model's 58% at its first checkpoint.

G Propositions and Proofs

Notation. Let $\delta \in [0,1]$ denote the detection probability and π denote the prior. Propositions 2 and 3 concern $\delta < 1$, since at $\delta = 1$ the $1-\delta$ rule and the posterior are 0 together, and the numerical confirmation in the code base is over the interior of $(0,1)$. After a failed, partial search the posterior of existence is $p^*(\delta,\pi) = \pi(1-\delta)/(1-\pi\delta)$ (Eq. (1)). Scoring uses the range of priors $\{p^*(\delta,\pi) : \pi \in [0.1, 0.9]\}$, instead of assuming any fixed prior rate. We make use of the two properties: p^* is strictly increasing in π and strictly decreasing in δ.

Proposition 1 (No scalar-only posterior recovery)**.** *If a post-hoc recalibration g depends only on the reported scalar q, and a model reports $q = 1-\delta$ for any $\delta \in (0,1)$, then no $g(q)$ equals $p^*(\delta,\pi)$, Eq. (1), for all $\pi \in (0,1)$. With the base rate as a second input, $h(q,\pi) = \pi q/(1-\pi+\pi q)$ recovers the posterior.*

Proof of Proposition 1. Take any g of the report $q = 1-\delta$ only. Since p^* is strictly increasing in π, two priors need two outputs and $g(q)$ is one number, so g misses at least one. Conversely h is Bayes' rule with q as the miss likelihood, and $h(1-\delta,\pi) = p^*(\delta,\pi)$. □

The identity $g(q) = q$ is right at one prior for each coverage, where $\pi(1+\delta) = 1$, and the argument applies to any δ-only response.

The stated base rate intervention of Sec. 3.1 tests the converse case empirically.

Corollary 1 (Irreducible error from an omitted prior)**.** *Fix any $\delta \in (0,1)$ and prior interval $[\pi_L, \pi_U] \subset (0,1)$. Let $p_L = p^*(\delta, \pi_L)$ and $p_U = p^*(\delta, \pi_U)$. Then, any scalar-only map g has a worst-case squared error*

$$\max_{\pi \in [\pi_L, \pi_U]} \big(g(1-\delta) - p^*(\delta, \pi)\big)^2 \;\geq\; \left(\frac{p_U - p_L}{2}\right)^2,$$

with equality at the midpoint $g(1-\delta) = (p_L + p_U)/2$.

In this appendix, model replies are scored against the prior interval $[\pi_L, \pi_U] = [0.1, 0.9]$, a range of plausible base rates and not a single prior; the rule assignment uses the posterior at $\pi = 0.5$ instead. Informally, one number cannot be close to both ends of the posterior's range over those priors, and the midpoint attains the bound, so it is sharp. The bound extends continuously to the zero-coverage endpoint $\delta = 0$. The companion bound below replaces the worst case by an average.

Proof. Since $p^*(\delta, \cdot)$ is continuous and strictly increasing, its range over the interval is $[p_L, p_U]$, so with $y = g(1-\delta)$ the worst case is at least $\max\{(y - p_L)^2, (y - p_U)^2\}$. By the triangle inequality $|y - p_L| + |p_U - y| \geq p_U - p_L$, the larger distance is at least $(p_U - p_L)/2$, with equality only at the midpoint. At $\delta = 0$, $p^*(0, \pi) = \pi$, so $p_L = 0.1$, $p_U = 0.9$, and the bound is $0.4^2 = 0.16$. □

Corollary 2 (Average-case error under a spread of base rates)**.** *Take any $\delta \in (0,1)$ and let the base rate π be drawn from any distribution over $(0,1)$ other than a point mass. Then any map g with only the report as input satisfies*

$$\mathbb{E}\big[(g(1-\delta) - p^*(\delta, \pi))^2\big] \;\geq\; \mathrm{Var}\big[p^*(\delta, \pi)\big] \;>\; 0,$$

with equality when $g(1-\delta)$ equals the mean of $p^*(\delta, \pi)$ under that distribution.

Proof. With $y = g(1-\delta)$ and $X = p^*(\delta, \pi)$, $\mathbb{E}[(y - X)^2] = (y - \mathbb{E}[X])^2 + \mathrm{Var}[X]$, with equality at $y = \mathbb{E}[X]$, and $\mathrm{Var}[X] > 0$ because $p^*(\delta, \cdot)$ is strictly increasing in π. □

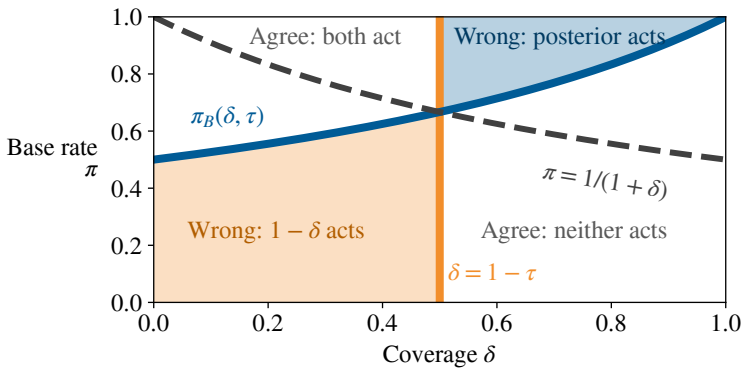


Figure 15: **Decision disagreement under a threshold** $\tau = 0.5$ (Prop. 2). The $1-\delta$ rule fires left of the vertical line $\delta = 1 - \tau$; the posterior fires above the curve $\pi_B(\delta, \tau) = 1/(2-\delta)$. Over the shaded regions the two induce different actions (orange: $1-\delta$ fires where the posterior would not; blue: the reverse). Below the dashed curve $\pi = 1/(1+\delta)$, $1-\delta$ overstates existence.

Proposition 2 (Decision-error region)**.** Assumptions. The downstream component fires if and only if the reported probability reaches a threshold $\tau \in (0,1)$. *The $1-\delta$ response and the posterior $p^*(\delta, \pi)$ induce different actions exactly between two boundaries. The $1-\delta$ response's action reverses on the vertical line $\delta = 1 - \tau$ and the posterior's reverses on the curve*

$$\pi_B(\delta, \tau) \;=\; \frac{\tau}{\tau + (1-\tau)(1-\delta)}. \tag{2}$$

Both disagreement regions are nonempty at each decision threshold τ. Call $1/(1+\delta)$ the crossing prior*, the single base rate at which $1-\delta$ equals the posterior (Sec. 2). The* direction *of the error follows from it: $1-\delta$ exceeds $p^*(\delta, \pi)$, overstating existence, at every true base rate below it. (This is a fact about the two formulas, not a claim that any model holds a prior; the main text's implicit-prior controls concern the latter.)*

Proof of Proposition 2. The $1-\delta$ response fires when $\delta \le 1-\tau$. For the posterior, $p^*(\delta,\pi) \ge \tau$ rearranges to $\pi \ge \pi_B(\delta,\tau)$, which rises strictly from τ at $\delta = 0$ to 1 at $\delta = 1$. The region where only $1-\delta$ fires contains $(0, 1-\tau) \times (0, \tau)$ because $\pi_B \ge \tau$, and the reverse region is nonempty at every $\delta > 1-\tau$ because $\pi_B < 1$. Since $(1-\delta) - p^*(\delta,\pi) = (1-\delta)\,(1-\pi(1+\delta))/(1-\pi\delta)$ with positive first factor and denominator at $\delta < 1$, $1-\delta$ exceeds the posterior exactly when $\pi < 1/(1+\delta)$, and at $\delta = 1-\tau$ both π_B and $1/(1+\delta)$ equal $1/(2-\tau)$, where the three boundaries meet (Fig. 15). □

Proposition 3 (Positive regret under proper scoring)**.** *Assume some fixed prior $\pi \in (0,1)$. At the endpoints p^* is constant in δ and a matching constant rule is correct. Responses are scored by a strictly proper scoring rule against the existence outcome, so the expected score is uniquely optimized by the true posterior (Gneiting & Raftery, 2007).*

Claims. *Every fixed rule, and $1-\delta$ in particular, pays positive regret except at a single crossing coverage. Regret here is the expected score's shortfall against the true posterior.*

- *(a) Any δ-independent rule has strictly positive regret except on a measure-zero set of δ.*
- *(b) At $\delta < 1$ the $1-\delta$ rule has strictly positive regret except at the single crossing $\delta = (1-\pi)/\pi$, which lies in $(0,1)$ only for $\pi > 1/2$ and is the crossing-prior condition $\pi = 1/(1+\delta)$ of Prop. 2.*
- *(c) A response that switches rules has the weighted average of their regrets.*

Closed form under the Brier score. *With $S(f,y) = (f-y)^2$, the regret of $f(\delta) = 1-\delta$ is $r(\delta,\pi) = \big((1-\delta)((1-\pi)-\pi\delta)/(1-\pi\delta)\big)^2$, whose value at the zero-coverage endpoint is $(1-\pi)^2$.*

Proof. Let $\bar{S}(f,p) = p\,S(f,1) + (1-p)\,S(f,0)$ and let the regret be $r = \bar{S}(f,p^*) - \bar{S}(p^*,p^*)$, nonnegative and zero only at $f = p^*$ by strict propriety, so each claim reduces to where the rule meets p^*. A constant meets $p^*(\cdot,\pi)$ at most once because p^* is strictly decreasing in δ; the $1-\delta$ rule meets it only where $\pi(1+\delta) = 1$, by the factoring in the proof of Prop. 2; and a response that switches rules inherits a weighted average of their regrets. Under the Brier score $r = (f-p^*)^2$, which at $f = 1-\delta$ gives the stated closed form and $(1-\pi)^2$ at $\delta = 0$. □

Lemma 1 (Within-group identification)**.** Assumptions. Two detection probabilities $\delta_1 < \delta_2$ presented within a coverage level that holds everything else fixed (the controlled within-group manipulation of Sec. 3, whose perturbation controls are the empirical support that nothing else moved). *The within-group response differences of the three candidate rules are strictly ordered:*

$$\underbrace{-(\delta_2-\delta_1)}_{1-\delta \text{ rule}} \;<\; \underbrace{p^*(\delta_2,\tfrac{1}{2}) - p^*(\delta_1,\tfrac{1}{2})}_{0.5\text{-prior posterior}} \;<\; \underbrace{0}_{\text{any count-based rule}},$$

so a single measured difference separates the three non-nested candidates. The identification is relative to this maintained candidate family (Lewbel, 2019).

Proof. A count-based rule depends on the searched count, constant within the group, so its difference is 0, and the $1-\delta$ rule's difference is $-(\delta_2-\delta_1)$. At $\pi = \frac{1}{2}$, $p^*(\delta,\frac{1}{2}) = (1-\delta)/(2-\delta)$ with derivative $-1/(2-\delta)^2 \in (-1,-\frac{1}{4})$ on $(0,1)$, so by the mean value theorem the posterior difference is $-(\delta_2-\delta_1)/(2-\varepsilon)^2$ for some $\varepsilon \in (\delta_1,\delta_2)$, strictly between $-(\delta_2-\delta_1)$ and 0. The margin between the two decreasing rules, $\big(1-1/(2-\varepsilon)^2\big)(\delta_2-\delta_1)$, shrinks toward zero as $\varepsilon \to 1$, so identification relies on interior coverage levels. □

H THE DEPLOYMENT SIMULATION

This part runs one record-keeping pipeline over the models' replies and counts the two errors that a threshold on a substituted probability produces. It carries the 31.0% of decisions suppressed at a base rate of 0.01 that Sec. 3.3 quotes (Table 26), together with the 34.7-point tie effect (Table 27), the 88% decline share (Table 28), and the 544 to 594 suppressions per $1{,}000$ decisions on the software-testing replies (Table 30).

H.1 THE PIPELINE AND ITS TWO ERRORS

A skip on an absent record is a silent loss; a create on a present record is a duplicate that can be found later. The pipeline asks "does a record exist? if judged missing, create it." After a failed search it creates the record unless the reported probability is 0.5 or more; a reply exactly at 0.5 skips.

It runs on two populations. The benchmark's own episodes are the answers of ten models to episodes from clinical records and enterprise documents, mixed at a base rate the pipeline is told and no prompt contains: each block of Table 26 holds $1{,}280$ episodes over four coverage levels, the clinical blocks at 0.89 split into the shallowest level ($n = 320$) and the three deeper ones ($n = 960$). The rates 0.01 and 0.30 are set for the evaluation; the clinical 0.89 is measured. The bands in Fig. 7 span 20 random draws of the episodes. Three clinical base rates appear in the paper: 0.89, measured on the emergency-department medication tables, is the one used here and in Table 1; 0.72 is the hospital medication tables' measured rate (App. B.2); and 0.6325 is the corpus's fitted $\hat{\pi}$ that the assumed-rate rows of Table 26 use. The reanalysis arrival stream is 90 days of reconciliation arrivals whose decisions are drawn from the measured per-(model, δ) value lists, with coverage 0 with probability 0.2, 0.32 with probability 0.5 and 1 with probability 0.3: $1{,}543$ decisions at $\pi = 0.3$ and $1{,}389$ at $\pi = 0.5$. The other tables of this part count this stream, per release.

H.2 THRESHOLDS AND BASE RATES

At $\pi = 0.3$ and $\tau = 0.5$ the correct posterior never crosses the threshold, so the oracle suppresses nothing and its only errors are the duplicates of creating every record (19.3% of decisions); thresholding $1 - \delta$ suppresses 56.9% of records and thresholding the 0.5-prior posterior 17.8%, almost all at zero coverage where its response is τ itself (Table 23). At $\pi = 0.89$ the trade reverses: even a deep failed search leaves the correct posterior above τ on a mostly-present population, so the true-rate row books suppressions where the raw number does not (20.6% against 3.1% of decisions on the deeper searches, Table 26), and the threshold, not the posterior, is the wrong instrument there.

Under a δ-dependent two-threshold rule, the band/certify rule that Table 23 defines, a skip cannot occur by construction; the table prices that safety in review volume, 76% of decisions at $\pi = 0.3$ and 82% at $\pi = 0.5$. Table 24 repeats the reanalysis at five thresholds. Below a base rate of 0.05 the two rules sit at their suppression maxima while the oracle's duplicates fall from 19.3% to 0.6% and 0.2% at 0.01 and 0.001 (Fig. 7, Table 25).

H.3 TEN METHODS ON THE BENCHMARK'S EPISODES

The episode stream mixes present and absent records at each coverage level, 160 and 160 per level per dataset. Among the present records, 360 of 640 in the clinical data and 298 of 640 in the enterprise data were found by the search, mostly at the two deeper levels; the model's mean report on those is 0.87 and 0.84, against 0.42 on absent records. Within a level, the report separates present from absent with an AUC of 0.58, 0.58, 0.88 and 0.85 (clinical) and 0.52, 0.64, 0.71 and 0.89 (enterprise) when the found records are included, and 0.46 to 0.54 when only records the search did not find are kept. Rows that read the model's number therefore gain from the found records, which a rule that reads only the coverage cannot see, and the coverage-only posterior is the correct posterior for the searches that found nothing. Platt and margin are refitted on the pooled levels in the base-rate 0.89 column, so they need not equal the average of the two halves below.

Table 26 splits every cell of Table 1 into suppressed, duplicated and reviewed decisions. *The model's number* thresholds the reported probability as it stands. *Posterior, assumed rate* and *Posterior, true rate* re-read that number as the miss likelihood and combine it with a base rate, the corpus's fitted rate (0.6325 clinical, 0.5267 enterprise) or the block's own. *Log only* computes the same posterior from the coverage in the search log and reads no reply; at the true rate it is the correct posterior. *Platt-scaled number* is a logistic rescaling of the reported number fitted on one half of the cell's own outcomes and applied to the other, an oracle because it sees the outcomes. *Conformal* calibrates an abstention threshold on one half so that the error among non-abstained decisions is at most α; refused means no threshold on the fitting half reaches α. *Margin abstention* is Chow's reject option, abstaining within a margin of 0.5 chosen on the calibration half for 80% coverage. *Create every record* and *Review everything* read nothing.

Table 23: **Each scalar rule pays in silent suppressions or duplicates; the band/certify rule pays in review.** Every number is a % of decisions, at $\tau = 0.5$, on one reanalysed population of 90 days of reconciliation arrivals under declared base rates $\pi = 0.3$ (1,543 decisions) and $\pi = 0.5$ (1,389 decisions). Scalar rows skip when the stated number reaches τ; the Bayes oracle applies that rule to the correct posterior; the band/certify rule creates when the whole band lies below τ, skips when it lies above, and otherwise reviews.

	$\pi = 0.3$ ($n = 1{,}543$)			$\pi = 0.5$ ($n = 1{,}389$)		
Rule at $\tau = 0.5$	Missed	Duplicated	Reviewed	Missed	Duplicated	Reviewed
Scalar rule, GPT-5.5	56.9	0.0	0%	47.6	0.0	0%
Scalar rule, GPT-5.4	55.0	0.5	0%	45.6	0.7	0%
Scalar rule, Opus 4.8	17.8	11.1	0%	14.9	20.2	0%
Scalar rule, Sonnet 5	47.8	2.3	0%	40.9	3.7	0%
Scalar rule, Haiku 4.5	56.6	0.1	0%	47.0	0.5	0%
Scalar rule, GPT-5.4-nano	15.9	13.0	0%	15.2	24.7	0%
Bayes oracle	0.0	19.3	0%	14.8	20.2	0%
Band/certify rule	0.0	0.0	76%	0.0	0.0	82%

One population is drawn once and every rule decides the same arrivals, so a difference between rows is not sampling noise. Each arrival's coverage is 0 with probability 0.2, 0.32 with probability 0.5, 1 with probability 0.3, and model answers are drawn from the measured value lists per model and level. A reply exactly at τ skips.

Table 24: **The reanalysed error rates across thresholds.** Each number is a % of reanalysed decisions, when each rule's stated number is thresholded at τ, on the same populations as Table 23. At these base rates moving τ exchanges one error for the other without removing either, the correct posterior included, and below 0.05 that stops (Table 25): at $\pi = 0.3$ the oracle suppresses nothing above $\tau = 0.3$ but duplicates every existing record, and at $\pi = 0.5$ it suppresses everything below $\tau = 0.5$.

Rule	π	Error	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$
$1 - \delta$ rule (GPT-5.5)	0.3	missed	56.9	56.9	56.9	54.8	15.6
		duplicated	0.0	0.0	0.0	1.2	12.3
0.5-prior posterior (Opus 4.8)	0.3	missed	56.6	37.3	17.8	0.1	0.0
		duplicated	0.1	5.2	11.1	19.3	19.3
Bayes oracle	0.3	missed	17.6	0.0	0.0	0.0	0.0
		duplicated	11.1	19.3	19.3	19.3	19.3
$1 - \delta$ rule (GPT-5.5)	0.5	missed	47.6	47.6	47.6	45.9	13.2
		duplicated	0.0	0.0	0.0	1.3	21.5
0.5-prior posterior (Opus 4.8)	0.5	missed	47.1	31.2	14.9	0.1	0.0
		duplicated	0.2	9.4	20.2	34.5	34.5
Bayes oracle	0.5	missed	47.6	47.6	14.8	0.0	0.0
		duplicated	0.0	0.0	20.2	34.5	34.5

The first five rows are our constructions; Platt scaling, conformal abstention at two levels and margin abstention are the published methods cited in Table 1, run on the same population. The conformal rule is refused at two of the five settings of Table 1 and sends 84.8%, 87.8% and 83.6% of decisions to review at the other three.

Table 25: **The same reanalysis at base rates of** 0.01 **and** 0.001**.** Each number is a % of decisions, missed or duplicated, at one seed, the format of Table 24. The oracle suppresses nothing at any τ, so moving the threshold gains nothing here, and the two rules lie at their maxima. Each rule's row is the one model standing for it at a single seed, so it can lie above the population-level maxima of Fig. 7.

Rule	π	Error	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$
$1 - \delta$ rule (GPT-5.5)	0.01	missed	71.5	71.5	71.5	69.2	18.4
		duplicated	0.0	0.0	0.0	0.0	0.4
0.5-prior posterior (Opus 4.8)	0.01	missed	71.1	45.9	20.8	0.1	0.0
		duplicated	0.0	0.2	0.4	0.6	0.6
Bayes oracle	0.01	missed	0.0	0.0	0.0	0.0	0.0
		duplicated	0.6	0.6	0.6	0.6	0.6
$1 - \delta$ rule (GPT-5.5)	0.001	missed	69.6	69.6	69.6	67.1	17.7
		duplicated	0.0	0.0	0.0	0.0	0.1
0.5-prior posterior (Opus 4.8)	0.001	missed	69.1	44.6	20.1	0.0	0.0
		duplicated	0.0	0.1	0.1	0.2	0.2
Bayes oracle	0.001	missed	0.0	0.0	0.0	0.0	0.0
		duplicated	0.2	0.2	0.2	0.2	0.2

Counts are PER RELEASE on the reanalysis arrival stream. Table 1 pools the schema cohort's ten over the benchmark's episodes, hence its smaller counts.

H.4 TIES

At $\tau = 0.5$ the choice of $\geq$ versus strict $>$ decides between total suppression and none for a model whose answer equals 0.5; the main text skips a reply exactly at τ. Table 27 recomputes Fig. 16 under both tie rules on the same population. A reply of exactly τ is common, so the rules differ at every base rate, by up to 4.5 points for $1 - \delta$ and 34.7 for the 0.5-prior posterior.

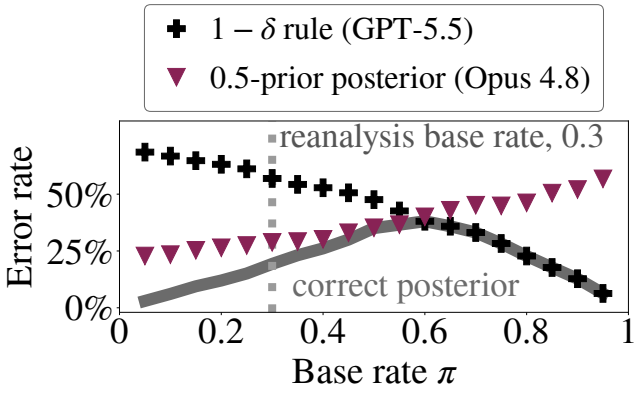


Figure 16: **Thresholding** $1 - \delta$ **suppresses records at low base rates, and thresholding the posterior at a prior of** 0.5 **duplicates them at high ones.** Two models, GPT-5.5 ($1 - \delta$ rule) and Opus 4.8 (posterior rule), at coverages 0, 0.32 and 1: decisions suppressed or duplicated at a threshold of 0.5. At these coverages the $1 - \delta$ rule makes the same decisions as the correct posterior above a base rate of 0.60. A reply exactly at 0.5 skips. That choice is not cosmetic: it moves the 0.5-prior posterior rule by up to 34.7 points (Table 27).

H.5 THE DECLINE CHANNEL

A review channel removes the silent error only for the models that use it. Tables 28 and 29 run the pipeline on every model's responses to the decline-permitted question, a decline sent to review and a

Table 26: **Every column behind Table 1.** Every number is a % of decisions. Wrong is missed plus duplicated, rounded once, so it can differ by one in the last place. The assumed rate is each corpus's fitted $\hat{\pi}$, 0.6325 clinical and 0.5267 enterprise, and the true rate is the block's own. Four rows are Table 1's published methods, with α 0.05 added, and *Create every record* and *Review everything* read nothing. Refused means no threshold on the fitting half reaches α, a finding and not a missing number. Episodes are weighted as Eq. 1 gives among failed searches at each coverage. The episode stream mixes present and absent records at each coverage level; about half of the present records overall, and most at the deeper levels, were found by the search, and rows that read the model's number gain from those, since on searches that found nothing the number carries no information beyond the coverage (App. H). Platt and margin are refitted on the pooled levels in the base-rate 0.89 column, so they need not equal the average of the two clinical 0.89 blocks here.

Method	Missed	Duplicated	Review	Wrong
clinical, base rate 0.01*, suppression side, L1–L4,* $n = 1{,}280$				
The model's number	31.0	0.2	0.8	31.2
Posterior, assumed rate	29.4	0.2	0.8	29.6
Posterior, true rate	0.0	0.5	0.8	0.5
Log only, assumed rate	49.6	0.2	0.0	49.8
Log only, true rate	0.0	0.6	0.0	0.6
Platt (Platt, 1999)	0.0	0.4	0.8	0.4
Conformal (Mohri & Hashimoto, 2024), α 0.10	0.5	0.1	84.8	0.5
Margin (Chow, 1970)	29.4	0.1	20.0	29.5
Create every record	0.0	0.6	0.0	0.6
Review everything	0.0	0.0	100.0	0.0
Conformal (Mohri & Hashimoto, 2024), α 0.05	0.5	0.1	84.8	0.5
clinical, base rate 0.30*, suppression side, L1–L4,* $n = 1{,}280$				
The model's number	23.3	5.4	0.9	28.7
Posterior, assumed rate	22.1	5.7	0.9	27.7
Posterior, true rate	0.0	18.3	0.9	18.3
Log only, assumed rate	37.2	5.8	0.0	43.0
Log only, true rate	0.0	18.6	0.0	18.6
Platt (Platt, 1999)	0.0	11.8	0.9	11.8
Conformal (Mohri & Hashimoto, 2024), α 0.10	refused	refused	refused	refused
Margin (Chow, 1970)	21.9	3.2	26.1	25.1
Create every record	0.0	18.6	0.0	18.6
Review everything	0.0	0.0	100.0	0.0
Conformal (Mohri & Hashimoto, 2024), α 0.05	refused	refused	refused	refused
clinical, base rate 0.89*, suppression side, L1,* $n = 320$				
The model's number	7.8	28.0	2.9	35.7
Posterior, assumed rate	7.6	28.0	2.9	35.6
Posterior, true rate	10.5	14.3	2.9	24.7
Log only, assumed rate	12.3	0.0	0.0	12.3
Log only, true rate	12.3	0.0	0.0	12.3
Platt (Platt, 1999)	12.1	0.0	2.9	12.1
Conformal (Mohri & Hashimoto, 2024), α 0.10	refused	refused	refused	refused
Margin (Chow, 1970)	7.5	14.8	20.0	22.3
Create every record	0.0	87.7	0.0	87.7
Review everything	0.0	0.0	100.0	0.0
Conformal (Mohri & Hashimoto, 2024), α 0.05	refused	refused	refused	refused
clinical, base rate 0.89*, duplicate side, L2–L4,* $n = 960$				

Table 27: **Fig. 16 under both tie rules at $\tau = 0.5$.** Wrong decisions as a % of decisions, on the same population. The main text skips a reply exactly at τ. The posterior equals τ at $\pi = 1/(2-\delta)$, so on this grid it can turn only at $\pi = 0.5$. The two rules threshold stored replies, where a reply of exactly τ is common, so they differ at every base rate here: up to 4.5 points for $1-\delta$ and 34.7 for the 0.5-prior posterior. The two numbers read tie-skips then tie-creates.

π	Tie	Correct posterior	$1-\delta$ (GPT-5.5)	0.5-prior (Opus 4.8)
0.05	skips	3.0	68.5	22.8
	creates	3.0	65.9	3.0
0.10	skips	6.1	66.7	23.4
	creates	6.1	64.8	6.1
0.15	skips	9.5	64.7	25.3
	creates	9.5	63.1	9.5
0.20	skips	12.0	63.1	26.4
	creates	12.0	61.5	11.9
0.25	skips	15.1	61.1	27.4
	creates	15.1	59.7	15.1
0.30	skips	19.3	56.9	28.8
	creates	19.3	56.0	19.3
0.35	skips	23.1	54.3	29.1
	creates	23.1	54.2	23.1
0.40	skips	26.1	52.8	30.1
	creates	26.1	52.5	26.1
0.45	skips	30.0	50.6	33.0
	creates	30.0	50.7	30.1
0.50	skips	35.1	47.6	35.2
	creates	34.5	47.2	34.6
0.55	skips	36.6	42.6	36.6
	creates	36.6	43.4	40.2
0.60	skips	38.1	38.1	40.0
	creates	38.1	38.7	45.9
0.65	skips	35.9	35.9	42.9
	creates	35.9	36.2	49.6
0.70	skips	33.1	33.1	45.1
	creates	33.1	34.0	54.0
0.75	skips	28.3	28.3	44.8
	creates	28.3	30.1	59.8
0.80	skips	22.8	22.8	46.2
	creates	22.8	24.3	67.2
0.85	skips	17.7	17.7	50.4
	creates	17.7	20.7	75.0
0.90	skips	12.8	12.8	52.0
	creates	12.8	16.0	81.6
0.95	skips	6.3	6.3	56.7
	creates	6.3	10.8	91.4

probability thresholded as before, with arrivals at the four partial-coverage levels the decline survey scores. Eight of the 40 releases the reanalysis reaches, those declining in 88% or more of required cases, produce no suppression; every release below that books some. Of the 13 models that decline on at least 60% of partial searches, 12 suppress silently on fewer than 10% of decisions at $\pi = 0.3$ while sending at least 60% of decisions to review; the exception, GPT-5-nano, declines least of the 13 at 61%. Where both channels can be operated, neither dominates on both error counts in 69 of 81 twinned pairs over the two base rates (the decline channel dominates in 1, the required probability in 11). A permitted decline cannot help once a number is required.

A certify-or-decline output, which lets the model escalate a record instead of scoring it, was run on six models in the same pipeline: it gives zero suppressions and zero duplicates for five of them by escalating every record, and fails for GPT-5.4-nano, which certifies about a third of the missed records absent. The per-row table of that comparison is not reproduced here; its counts are in the released code.

H.6 The software-testing replay

The software-testing replies cost the same in the same pipeline. Run on that setting's responses at the same base rate and threshold, it gives Table 30, with the benchmark's reanalysis of the same release beside it where one is on record. The three releases measured in both settings suppress 544, 569 and 594 records per 1,000 decisions on the software-testing replies; all eleven lie between 479 and 594, and the correct posterior suppresses none.

H.7 Proper scoring

Given a stated prior, most models score near zero expected Brier regret against the stated-prior posterior on the shared priors ($\pi \in \{0.1, 0.3, 0.7\}$), computed without outcome sampling because the search model is known. Five score zero to three places, against 0.185 for a $1 - \delta$ response and 0.061 for a 0.5-prior posterior (Fig. 17).

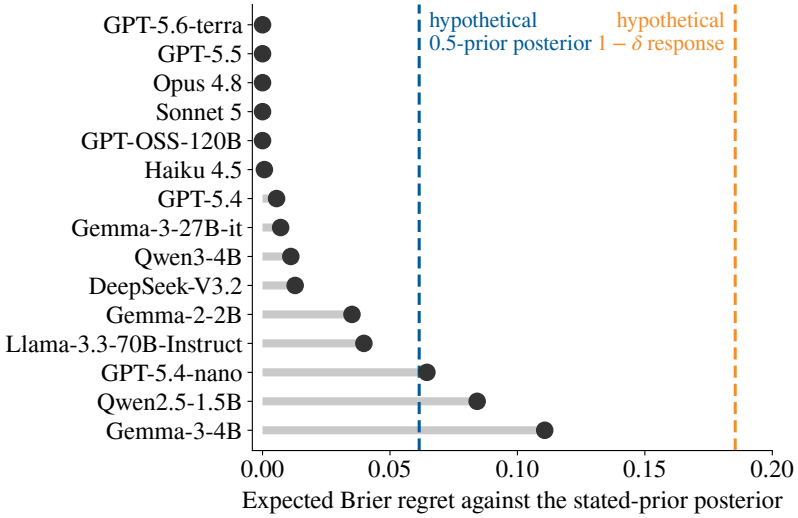


Figure 17: **Most measured models beat both fixed rules under proper scoring.** Expected Brier regret of each model's stated-prior responses against the stated-prior posterior, over the shared priors ($\pi \in \{0.1, 0.3, 0.7\}$). Dashed lines score a fixed $1 - \delta$ response (0.185) and a 0.5-prior posterior response (0.061). Three releases do worse than the 0.5-prior posterior rule: Gemma-3-4B, Qwen2.5-1.5B and GPT-5.4-nano; Gemma-2-2B, smaller than the first, beats it.

Table 28: **The record-keeping reanalysis with a decline channel, at a base rate of** 0.3**.** Sec. 3.3's pipeline reanalysed on each model's decline-permitted responses at the four partial levels, zero-coverage arrivals excluded, coverage disclosed, thresholded at $\tau = 0.5$, declines to review. Every number is a % of decisions, except Declined, a share of the model's own responses from the survey (Table 13). The required suppression and duplication counts are issued with the code.

Release	Declined	Suppressed	Duplicated	Review
Opus 4.8	100%	0.0%	0.0%	100%
Gemini 3.1 Pro	100%	0.0%	0.0%	100%
GPT-5.5	100%	0.0%	0.0%	100%
GPT-5.6-sol	100%	0.0%	0.0%	100%
Grok 4.20 Reasoning	100%	0.0%	0.0%	100%
Fable 5.1	97%	0.0%	0.7%	97%
Fable 5	88%	0.0%	2.8%	87%
o1	88%	0.0%	1.3%	89%
Qwen3-8B	72%	1.3%	4.3%	73%
GPT-5-mini	67%	7.8%	2.7%	69%
Qwen3-14B	64%	2.7%	3.4%	68%
Gemma-3-12B-it	62%	6.5%	4.3%	65%
GPT-5-nano	61%	11.5%	3.2%	63%
GPT-4 Turbo	45%	7.7%	5.3%	49%
o3-mini	44%	9.1%	5.3%	48%
GLM-5	44%	17.9%	4.0%	47%
GPT-4o-mini	38%	6.2%	8.3%	40%
Qwen3-32B-AWQ	34%	13.0%	6.5%	37%
Qwen3-32B	30%	16.7%	6.5%	32%
Nova Pro	30%	17.8%	6.4%	32%
Llama 4 Maverick	30%	7.5%	9.8%	32%
o4-mini	28%	27.8%	4.3%	29%
Qwen3-32B-FP8	27%	19.1%	6.4%	29%
DeepSeek-V3.2	25%	26.1%	5.2%	27%
Qwen2.5-72B-Instruct	22%	26.9%	5.7%	23%
GPT-4.1-nano	19%	3.2%	15.8%	19%
Haiku 4.5	17%	30.1%	5.5%	18%
Mistral Large 3	16%	16.2%	9.5%	17%
Sonnet 4.5	12%	30.2%	6.0%	14%
GPT-4o	12%	30.3%	6.2%	14%
GPT-5.2	12%	26.1%	7.3%	13%
o3	5%	35.1%	6.3%	5%
Gemma-2-27B-it	3%	37.7%	6.0%	3%
GPT-3.5 Turbo	2%	14.7%	17.3%	2%
GPT-4	2%	39.0%	6.0%	2%
GPT-5.4-mini	2%	32.6%	8.9%	2%
GPT-4.1	0%	36.3%	7.3%	0%
GPT-4.1-mini	0%	36.3%	7.2%	0%
GPT-5.4-nano	0%	22.7%	12.5%	0%
Qwen3-30B-A3B	0%	11.6%	16.9%	0%
Kimi K2.7 (8,000 tokens)	100%	0.0%	0.0%	100%
Sonnet 5 (8,000 tokens)	14%	28.0%	6.0%	17%
Oracle posterior	–	0.0%	20.3%	0%

Table 29: **The record-keeping reanalysis with a decline channel, at a base rate of** 0.5**.** Sec. 3.3's pipeline reanalysed on each model's decline-permitted responses at the four partial levels, zero-coverage arrivals excluded, coverage disclosed, thresholded at $\tau = 0.5$, declines to review. Every number is a % of decisions, except Declined, a share of the model's own responses from the survey (Table 13). The required suppression and duplication counts are issued with the code.

Release	Declined	Suppressed	Duplicated	Review
Opus 4.8	100%	0.0%	0.0%	100%
Gemini 3.1 Pro	100%	0.0%	0.0%	100%
GPT-5.5	100%	0.0%	0.0%	100%
GPT-5.6-sol	100%	0.0%	0.0%	100%
Grok 4.20 Reasoning	100%	0.0%	0.0%	100%
Fable 5.1	97%	0.0%	1.3%	97%
Fable 5	88%	0.0%	5.1%	87%
o1	88%	0.0%	2.7%	90%
Qwen3-8B	72%	1.3%	8.4%	74%
GPT-5-mini	67%	6.0%	5.0%	69%
Qwen3-14B	64%	1.9%	6.7%	70%
Gemma-3-12B-it	62%	5.5%	8.6%	64%
GPT-5-nano	61%	9.5%	5.9%	63%
GPT-4 Turbo	45%	6.0%	9.6%	51%
o3-mini	44%	7.0%	10.0%	50%
GLM-5	44%	13.9%	7.6%	47%
GPT-4o-mini	38%	5.0%	16.0%	41%
Qwen3-32B-AWQ	34%	9.0%	12.3%	39%
Qwen3-32B	30%	12.8%	11.9%	33%
Nova Pro	30%	13.8%	11.8%	33%
Llama 4 Maverick	30%	5.1%	19.0%	33%
o4-mini	28%	20.6%	8.3%	29%
Qwen3-32B-FP8	27%	14.2%	11.8%	30%
DeepSeek-V3.2	25%	19.8%	10.0%	27%
Qwen2.5-72B-Instruct	22%	20.7%	10.3%	24%
GPT-4.1-nano	19%	2.6%	30.3%	18%
Haiku 4.5	17%	22.8%	10.0%	19%
Mistral Large 3	16%	11.7%	18.3%	19%
Sonnet 4.5	12%	22.9%	11.2%	14%
GPT-4o	12%	22.9%	11.4%	14%
GPT-5.2	12%	18.8%	14.0%	14%
o3	5%	26.1%	11.7%	6%
Gemma-2-27B-it	3%	28.1%	11.0%	4%
GPT-3.5 Turbo	2%	12.0%	32.5%	2%
GPT-4	2%	29.3%	11.0%	2%
GPT-5.4-mini	2%	24.4%	16.9%	2%
GPT-4.1	0%	27.1%	13.4%	0%
GPT-4.1-mini	0%	27.5%	13.6%	0%
GPT-5.4-nano	0%	16.6%	23.8%	0%
Qwen3-30B-A3B	0%	8.7%	31.3%	0%
Kimi K2.7 (8,000 tokens)	100%	0.0%	0.0%	100%
Sonnet 5 (8,000 tokens)	14%	21.4%	11.6%	17%
Oracle posterior	–	0.0%	39.0%	0%

Table 30: **The software-testing responses cost the same in the same pipeline.** The BugSearch replies pushed through the record-keeping pipeline of Sec. 3.3 at the same base rate ($\pi = 0.3$), the same threshold ($\tau = 0.5$) and the same weights on pre-search arrivals ($\delta = 0.0$ w.p. 0.2, $\delta = 0.32$ w.p. 0.5, $\delta = 1.0$ w.p. 0.3), over the 1,543 of 1,800 arrivals with a failed search. Silent suppressions and duplicate records per 1,000 decisions. The NullSearch column is the same release's reanalysis in the search setting where one is on record.

Release	suppressions	duplicates	NullSearch supp.
GPT-5.5	569	0	569
Grok 4.20 Reasoning	569	0	–
Nova Pro	544	10	–
Haiku 4.5	544	3	566
Kimi K2.7	569	0	–
Sonnet 5	594	0	478
GLM-5	550	14	–
DeepSeek-V3.2	479	44	–
Mistral Large 3	529	21	–
Llama 4 Maverick	493	21	–
Gemini 3.1 Pro	518	15	–
Correct posterior (oracle)	0	193	0

I REPRODUCIBILITY

Seeds and replication. The environment is generated deterministically from a seed; main results use seed 3, and rule assignments are stable across the seeds collected except for GPT-4o, the closest to the 60% cutoff of the 20 releases collected at both seeds (the note at the end of this appendix gives the seeds and counts). The 1,441 August responses were kept cut to 300 characters, which touches part of one stated-prior case only. Every case collected from September 2026 on was fixed before its first call, and each quoted count comes from two independently written scorers. The generator, the scoring code and all raw outputs go with the paper as supplementary material and will be made public, with open-weight checkpoints fixed by snapshot hash. The external datasets remain under their own licenses, and no clinical record content is redistributed; App. B gives MIMIC's access terms.

Behind the model table. The provider counts within each rule group of Table 7, the releases with no parsed response, the replies lost to truncation and the supplied identifier of every model are recorded here, not under the table. Omitted, with no parsed response at the four partial-coverage levels: Kimi-K2.6 (256 queries); MiniMax-M3 (1 query); Nemotron 3 Ultra (128 queries); Opus 5 (Bedrock, 6 queries); Opus 5 (refuses the instrument on the direct channel). Sampling is each provider's default, with the temperature unset in these runs. The seeds vary the episodes, not the decoder. Of the 20 models collected at seeds 3 and 11, only GPT-4o is assigned a different rule. Of those 20, GPT-4o and GPT-4 are the two closest to the 60% cutoff, and at 11 GPT-4 keeps its rule by the minimum count, not a reply to spare. n counts polled responses; truncated or unparsable replies, excluded from every count: Kimi K2.7 21, GPT-5 15, GPT-5.5 10, Gemini 3.1 Pro 6, GPT-3.5 Turbo 4, Opus 4.8 3, Sonnet 5 1.