Why Hierarchical Structure Matters for Uncertainty Quantification in Agentic AI Evaluation

Lennart Luettgau [*1] **Harry Coppock** [*1] **Magda Dubois** [1]
Christopher Summerfield [1] **Cozmin Ududec** [1]

Abstract

Uncertainty quantification is critical for trustworthy AI evaluations, yet the statistical methods commonly used were designed for data-rich settings and largely ignore the hierarchical structure of evaluation data. Here, we compare four uncertainty quantification methods that differ in how much they account for evaluation data structure ("naive" confidence intervals, item-level bootstrap, cluster bootstrap, and hierarchical Bayesian GLMs), across QA benchmarks (use cases 1 and 2) and an agentic benchmark (use case 3), where we additionally assess their reliability in low-data regimes. Across both setups, methods that do not account for hierarchy produce biased performance estimates that may lead to spurious inferences of significant differences between models. The cluster bootstrap accounts for hierarchy but its point estimates shift substantially across data regimes, limiting its reliability. Hierarchical Bayesian GLMs produce less biased estimates in QA benchmarks and the most stable estimates in the agentic benchmark, with reasoning effect estimates $4\times$ less sensitive to data regime than the cluster bootstrap. These findings demonstrate that accounting for hierarchical data structure is essential for reliable uncertainty quantification in AI evaluation. Despite being well-established in the natural and social sciences, hierarchical GLMs remain largely absent from AI evaluation practice; to encourage their adoption, we implement all models in the open-source HiBayES package.

[*]Equal contribution [1]AI Security Institute, UK. Correspondence to: Lennart Luettgau <lennart.luettgau@dsit.gov.uk>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Evaluating AI systems, particularly agentic systems that autonomously use tools, execute multi-step plans, and interact with external environments, requires uncertainty estimates that can be trusted enough to make safety judgments and deployment decisions. When a single evaluation run on an agentic benchmark can cost hundreds of dollars in tokens (Kamradt, 2025) and take days to complete, researchers cannot afford to discover after collecting the full dataset that their early estimates were misleading. Yet current evaluation practices provide little basis for confidence: the dominant approach reports aggregate accuracy without adequate uncertainty bounds, largely following the "highest-number-is-best" heuristic (Miller, 2024), without accounting for the structure of benchmark data. A recent review of LLM benchmarks found widespread patterns in task design, measurement, and scoring that undermine the validity of evaluation claims, often related to a lack of appropriate statistical analysis of the data (Bean et al., 2025).

This matters because evaluation data is inherently nested. Item-level responses cluster within benchmarks, benchmarks cluster within capability domains, and evaluations are typically conducted to compare multiple models (Figure 1). Each level introduces its own source of variation, yet current approaches to evaluation statistics treat all items as independent ("flat models"), discarding important information about the data structure. Research in the natural and social sciences has consistently shown that such approaches produce overconfident estimates unlikely to generalize (McElreath, 2020). The problem is compounded in agentic settings, where additional structure arises from varying task difficulty, reasoning effort levels, and scaffold configurations, and where the economic cost of data collection makes low-data regimes the norm rather than the exception, challenging assumptions of commonly used statistical methods around data availability and convergence.

Recent proposals to improve AI evaluation statistics, including error bars, t-tests (Miller, 2024) or simple Bayesian models (Bowyer et al., 2025) represent meaningful advances but do not systematically address the full hierarchical nest-

ing of evaluation data and the adherent requirement for propagation of uncertainty estimates, nor do they provide head-to-head empirical comparisons between parametric and distribution-free alternatives or demonstrate stability of estimates across data regimes (see Section 2 for a detailed discussion).

In this paper, across three evaluation scenarios of increasing complexity, we systematically compare four distinct approaches to uncertainty quantification (naive CIs, item bootstrap, cluster boostrap and hierarchical Bayesian GLMs; cf. Table 1 for detail). Our key contributions are:

1. **In QA benchmarks (use cases 1–2)**, hierarchical Bayesian GLMs produce less biased overall performance estimates than methods ignoring hierarchy, and reveal that apparent model differences are driven by large N rather than meaningful effect sizes (Section 4.1).
2. **In an agentic benchmark (use case 3)**, hierarchical Bayesian GLMs produce the most stable estimates across data regimes, with reasoning effect estimates $4\times$ less sensitive to data regime than the cluster bootstrap, enabling earlier resource allocation decisions (Section 4.2).
3. **Cluster bootstrapping**, while hierarchy-aware and distribution-free, produces unstable estimates, limiting its reliability in both settings (Section 4.1, Section 4.2).

We do not claim methodological novelty in the statistical tools themselves, which are well-established in the natural and social sciences (Gelman et al., 2013; McElreath, 2020) but remain largely absent from AI evaluation practice. Our contribution is demonstrating empirically that adopting these tools changes substantive conclusions about model capabilities, and providing practical guidance for the AI evaluation community on when and why this matters. While our focus is on evaluation-time uncertainty, instead of deployment-monitoring, the two are deeply connected: if evaluation-based capability estimates are miscalibrated (as we show conventional methods produce) then downstream deployment decisions (e.g., whether a model is capable enough to be routed to a particular task class, or whether additional reasoning effort is cost-justified) that rely on those estimates inherit that miscalibration.

The stability advantage of hierarchical Bayesian GLMs is particularly consequential in low-data regimes. Partial pooling regularizes estimates even when per-cell data is scarce, producing point estimates that change minimally as additional data arrives. This contrasts with both naive CIs (which narrow substantially as data accumulates, suggesting poor early calibration) and cluster bootstrapping (which produces consistent interval widths but point estimates that shift markedly). For agentic evaluations where each epoch of data collection is expensive, the practical difference is between making reasonably reliable early decisions and having to wait for the full dataset.

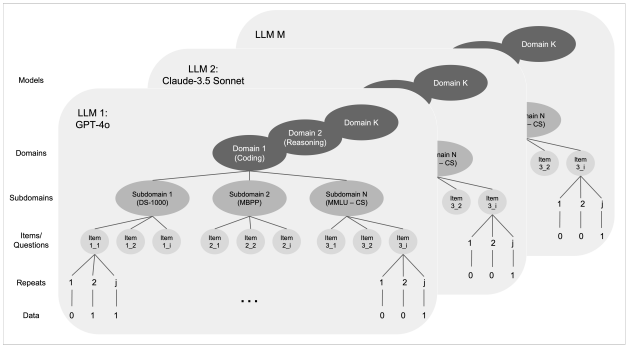

Figure 1. **Hierarchically nested structure of evaluation data.** Responses are nested within items, which are nested within benchmarks (subdomains), which belong to broader domains. Multiple models are typically evaluated on the same set of tasks.

To lower barriers for adoption of these methods, all models in this paper are implemented using the open-source HiBayES software package (Coppock et al., 2025), which provides accessible hierarchical Bayesian modeling for AI evaluation. Built on NumPyro (Phan et al., 2019), HiBayES reads standard evaluation log files (e.g., from the Inspect framework; UKGovernmentBEIS 2024), supports specification and fitting of multilevel GLMs via MCMC, and returns posterior summaries with uncertainty intervals and model comparison metrics, requiring no prior experience with Bayesian modeling. Full model specifications are provided in Appendix A.

2. Related work

AI evaluation practices. As LLMs and agentic systems have grown increasingly capable, the AI community has developed a wide range of benchmarks spanning coding (Yu et al., 2024; Lai et al., 2022), reasoning (Clark et al., 2019; Lai et al., 2017; Hendrycks et al., 2020; Shi et al., 2022), and agentic task completion (Mialon et al., 2023; Yang et al., 2023; Rein et al., 2025; Han et al., 2023). Despite the sophistication of these benchmarks, the statistical methods used to analyze their results remain rudimentary. The dominant practice follows the "highest-number-is-best" approach (Miller, 2024), reporting aggregate accuracy without uncertainty quantification. Leaderboards rank models by point estimates, and claims about model superiority or threshold-crossing are often made without formal statistical validation. Dror et al. (2018) documented widespread misuse or omission of statistical testing in NLP and proposed a practical protocol for significance test selection. Miller (2024) recommends adding error bars and t-tests to evaluation reports, representing a step beyond point estimates. However, t-tests assume approximately Gaussian data, re-

Table 1. Properties of the four uncertainty quantification methods compared in this paper. Use cases 1 and 2 are QA benchmarks; Use case 3 is an agentic benchmark (cf. Section 3.3 for more information). Legend: ✓ = yes; ∼ = partially (the cluster bootstrap is strongly influenced by the number and item imbalance of clusters considered), blank cell: no

Uncertainty quantification method	Tested on QA benchmark (use cases 1 and 2)	Tested on agentic benchmark (use case 3)	Accounts for hierarchical structure	Partial Pooling/ Regularization	Reliable in low-data regime
Naive CI	✓	✓			
Item bootstrap	✓				
Cluster bootstrap	✓	✓	✓	∼	
Hierarchical GLMs	✓	✓	✓	✓	✓

quire large samples to satisfy asymptotic assumptions, and cannot account for hierarchical nesting of data.

Bayesian modeling for AI evaluation. Multilevel (hierarchical) models have a long history in the natural and social sciences for analyzing nested data structures (Gelman et al., 2013; McElreath, 2020). These models allow parameters to vary across levels of a data hierarchy while partially pooling estimates toward group means, producing regularized estimates that are less prone to overfitting and more likely to generalize. Bayesian posterior distributions enable principled uncertainty quantification through credible intervals and highest posterior density intervals (HPDIs), without relying on asymptotic assumptions. Within AI evaluation, several recent papers have begun to apply Bayesian methods: Bowyer et al. (2025) propose simple Bayesian models (e.g., a Binomial model for binary outcomes) as alternatives to Frequentist confidence intervals. Longjohn et al. (2025) develop a Bayesian approach for quantifying uncertainty in binary evaluation metrics induced by stochastic text generation, demonstrating its utility for evaluating refusal rates and pairwise preferences. Llewellyn et al. (2025) propose a Bayesian hierarchical model with embedding-space clustering for evaluating LLM vulnerabilities to prompt injection, showing that accounting for output variability can yield better calibrated findings.

Distribution-free and conformal methods. Distribution-free methods, including the bootstrap (Efron, 1979), cluster bootstrap (Cameron et al., 2008), conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2021; Romano et al., 2019), and prediction-powered inference (Angelopoulos et al., 2023), provide coverage guarantees without parametric assumptions and have attracted growing interest in machine learning (Barber et al., 2023). However, these methods do not naturally incorporate hierarchical structure or partial pooling: when evaluation data is nested, they must either treat all items as exchangeable (ignoring clustering) or resample at the cluster level, which becomes unstable when clusters are few, a common situation in AI evaluation. We view hierarchical Bayesian models and distribution-free methods as complementary rather than competing. Recent work on group-conditional conformal prediction (Barber et al., 2023) and conformalized Bayes (Angelopoulos and Bates, 2021) suggests natural connecting points: posterior predictive distributions from hierarchical GLMs could serve as nonconformity scores, yielding prediction sets that respect evaluation hierarchy while retaining finite-sample coverage guarantees. Exploring such hybrid approaches is beyond the scope of this work but represents a promising direction.

3. Methods

AI evaluation data across QA evals and agentic setups is typically hierarchically structured: binary item-level outcomes (correct/incorrect) are nested within benchmarks (evaluations), which are nested within capability domains, with evaluations often conducted across multiple models (Figure 1)

We compare four approaches to uncertainty quantification (cf. Table 1). **Naive CIs** pool all items, ignoring clustering, with confidence interval $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N}$. **Item bootstrap** resamples items with replacement from the pooled set (2,000 replicates), also ignoring clustering. **Cluster bootstrap** resamples benchmarks with replacement, then items within each resampled benchmark (2,000 replicates), respecting clustering structure. **Hierarchical Bayesian GLMs** with Binomial or Beta-Binomial likelihoods and partial pooling across hierarchy levels (McElreath, 2023). assume that parameters at lower levels are drawn from higher-level distributions, enabling information sharing across clusters. Posteriors are estimated via NUTS (Hoffman and Gelman, 2011) in NumPyro (Phan et al., 2019) with 2,000 warmup and 2,000 inference samples across 4 chains; convergence diagnostics are in Appendix A.5. Uncertainty is quantified via 95% HPDIs. Formal comparison between GLMs that differ in their structural assumptions enables a comparison of competing hypotheses about the data-generating process (e.g., whether reasoning effort affects performance, or whether task-level overdispersion must be modeled), we use leave-one-out cross-validation (LOO-CV) to estimate out-of-sample pre-

dictive accuracy. Full specifications are in Appendix A.

We note that Frequentist mixed-effects models also accommodate hierarchical structure through random effects and partial pooling, and would likely produce point estimates similar to our hierarchical Bayesian GLMs for Use Cases 1 and 2. Our set of methodological approaches was chosen to represent the methods currently used or recently recommended in the AI evaluation literature (Dror et al., 2018; Miller, 2024).

3.1. Hierarchical Bayesian GLMs: a worked example

To illustrate the GLM approach, we describe the simplest of the models used in this paper: the Binomial GLM for use case 1. The models for use cases 2 and 3 extend the same structure with additional levels (models, tasks, reasoning effort) and, where appropriate, a Beta-Binomial likelihood to accommodate overdispersion; full specifications are provided in Appendix A.

Data is aggregated to the benchmark level: for each benchmark j within domain d, we observe Correct_j successes out of n_j trials. The model is:

$$\begin{aligned}
\text{Correct}_j &\sim \text{Binomial}(n_j, p_j) \\
\text{logit}(p_j) &= \text{domain}_{d[j]} \\
\text{domain}_d &\sim \text{Normal}(\bar{\mu}, \sigma_{\text{domain}}) \\
\bar{\mu} &= \mu_{\text{overall}} + \sigma_{\text{overall}} \cdot z_{\text{overall}} \\
\mu_{\text{overall}} &\sim \text{Normal}(0, 1) \\
\sigma_{\text{overall}} &\sim \text{HalfNormal}(0.5) \\
z_{\text{overall}} &\sim \text{Normal}(0, 1) \\
\sigma_{\text{domain}} &\sim \text{HalfNormal}(0.1).
\end{aligned} \tag{1}$$

The *Binomial likelihood* models benchmark-level success counts. The *hierarchical prior* on domain_d implements partial pooling: domain-level estimates are regularized toward $\bar{\mu}$, with pooling strength learned from the data. The overall performance estimate is the posterior of $\bar{\mu}$, which weights domains equally rather than being dominated by the one with most samples. We use a non-centered parameterization for sampling efficiency and mildly informative priors on the logit scale.

3.2. Estimation stability analysis

In use case 3, we assess the sensitivity of each method's estimates to data regime by comparing results obtained from 3 epochs (early, limited data) against the full dataset ($\sim$10 epochs). We evaluate two metrics: (1) *interval width reduction* — how much 95% intervals shrink as data accumulates (large reductions suggest the early intervals were poorly calibrated); and (2) *slope drift* — the absolute change in reasoning effect point estimates between data regimes (large drift indicates the early estimate was unreliable as a basis for decisions). We emphasize that this stability analysis is conducted on a single agentic benchmark (GAIA) and should be interpreted as illustrative rather than definitive. Replication across additional agentic benchmarks (e.g., SWE-bench) is needed to establish generalizability.

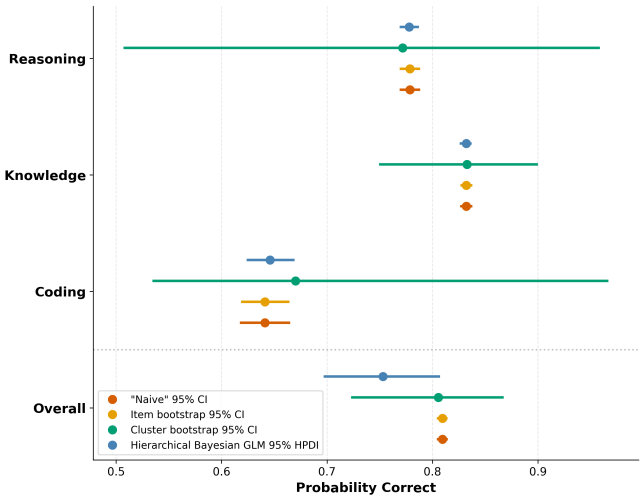

Figure 2. **Domain-level and overall performance estimates across four methods.** Each row shows the point estimate (dot) and 95% interval (horizontal line) for a given domain or overall performance on the probability scale. Methods compared: naive 95% CI, item bootstrap (resampling items), cluster bootstrap (resampling at the benchmark level), and hierarchical Bayesian GLM with partial pooling.

3.3. Datasets

Dataset 1 (use case 1). One LLM (Claude Sonnet-4) evaluated across 12 benchmarks spanning three domains: coding ($N = 1{,}728$), knowledge ($N = 25{,}035$), and reasoning ($N = 9{,}030$), with 4 benchmarks per domain.

Dataset 2 (use case 2). Four LLMs (GPT-4o, GPT-5.4, Claude Sonnet-4, Claude Sonnet-4-6) evaluated across the same domains and benchmarks as above.

Dataset 3 (use case 3). Four LLMs (o1, o3-mini, Claude-3.7-Sonnet, GPT-4.5-preview) evaluated on GAIA (Mialon et al., 2023), an agentic benchmark with 165 tasks at three difficulty levels. o1 and o3-mini were tested at four reasoning effort levels (none, low, intermediate, high); the remaining models were tested without reasoning only. Each task was repeated 10 times per condition, yielding 11,650 total observations distributed unevenly across models.

All evaluations were conducted using the open-source AI evaluation framework Inspect (UKGovernmentBEIS, 2024).

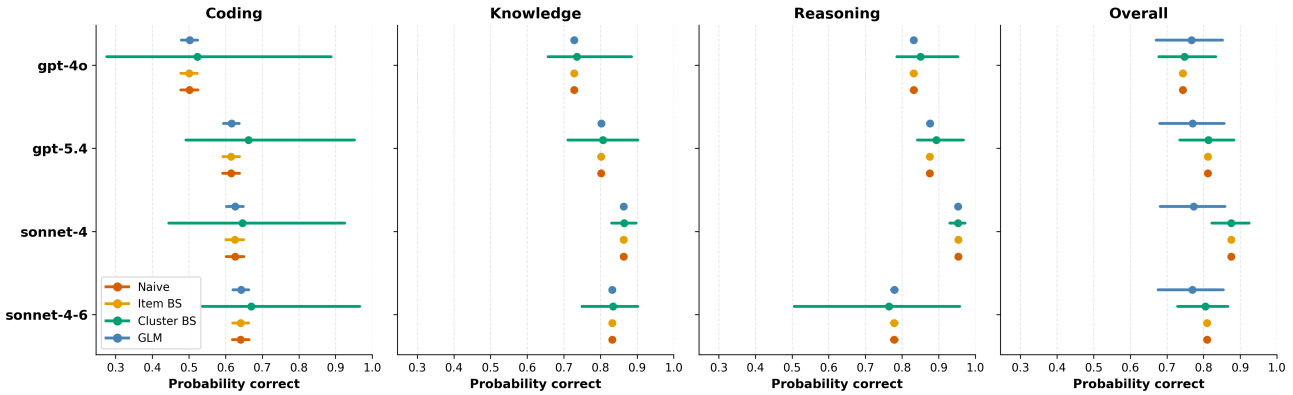

Figure 3. **Model-level and domain-level performance estimates across four methods.** Within each panel, rows correspond to the four evaluated LLMs (GPT-4o, GPT-5.4, Sonnet-4, Sonnet-4-6), with point estimates (dots) and 95% intervals (horizontal lines) on the probability scale. Methods compared: naive 95% CIs (orange), item bootstrap (yellow), cluster bootstrap (green), and hierarchical Bayesian GLM (blue).

4. Results

4.1. Use cases 1–2: QA benchmarks

Overall performance estimation (use case 1). A common target claim of evaluations is estimating a model's overall capability across several benchmarks. It is still common practice to report a single aggregate accuracy without quantifying uncertainty or accounting for how data is distributed across benchmarks, leaving open whether observed performance reflects overall capability or is heavily influenced by uneven sampling across evaluations. We evaluate one LLM across three domains with substantial item imbalances (coding: $N = 1{,}728$; knowledge: $N = 25{,}035$; reasoning: $N = 9{,}030$). Naive CIs and item bootstrap estimate overall performance at 0.824 (width 0.008), dominated by the knowledge domain which contributes 70% of items. Hierarchical Bayesian GLM estimates 0.762 (HPDI width: 0.120), placing the overall estimate centrally between domains (Figure 2). For a deployment threshold of 80%, naive methods suggest it is met, whereas Hierarchical Bayesian GLMs indicate it may not be. The cluster bootstrap produces unstable domain-level intervals (coding width: 0.436) due to only 4 benchmarks per domain. Item-level t-tests find all domain pairs significantly different ($t = 18$, $p < .001$); benchmark-level t-tests find none significant ($p = .483$). Hierarchical Bayesian GLMs resolve all comparisons with directly interpretable effect magnitudes.

Comparing overall performance of multiple models (use case 2). A second common scenario is comparing multiple LLMs across benchmarks to determine which model is better overall. Current practice typically compares models pairwise using t-tests or by inspecting non-overlapping confidence intervals, declaring any statistically significant difference as evidence that one model outperforms another. We compare four LLMs across the same benchmarks. Item-level t-tests declare nearly all pairs significantly different (e.g., $t = 41.9$, $p < .001$), driven by large N. The cluster bootstrap produces erratic point estimates and wide intervals crossing zero. Hierarchical Bayesian GLMs find all pairwise differences near zero (all HPDIs contain zero, posterior probabilities ~ 0.50), supporting no conclusion of overall differences (Figure 3). We verify this reflects genuine similarity rather than excessive shrinkage (Appendix A.6).

4.2. Use case 3: Agentic evaluation: Uncertainty estimation in low-data regimes

Agentic evaluations are costly: token costs can reach hundreds of dollars per task (Kamradt, 2025), take up substantial amounts of expert time in setup and may take hours to run to completion, creating low-data regimes where a central question is whether statistical estimates are trustworthy enough to guide resource allocation decisions.

We evaluate four LLMs (o1, o3-mini, Claude-3.7-Sonnet, and GPT-4.5-preview) on GAIA (Mialon et al., 2023), an agentic benchmark comprising 165 unique tasks at three difficulty levels (easy, mid, hard). To assess how reasoning effort affects performance, o1 and o3-mini were tested at four reasoning levels (none, low, intermediate, high), while Claude-3.7-Sonnet and GPT-4.5-preview were tested without reasoning only. Each task was repeated 10 times per condition, yielding 11,650 total observations distributed unevenly across models (o1: 4,632; o3-mini: 4,664; Claude-3.7-Sonnet: 1,565; GPT-4.5-preview: 789).

Investigating per-task success rates across models, we observe an asymmetric bimodal distribution: for the majority of tasks, models either pass all attempts or fail all attempts, with intermediate success rates being rare. This pattern is characteristic of agentic evaluations: tasks tend to be either saturated (trivially solved) or too difficult, providing limited

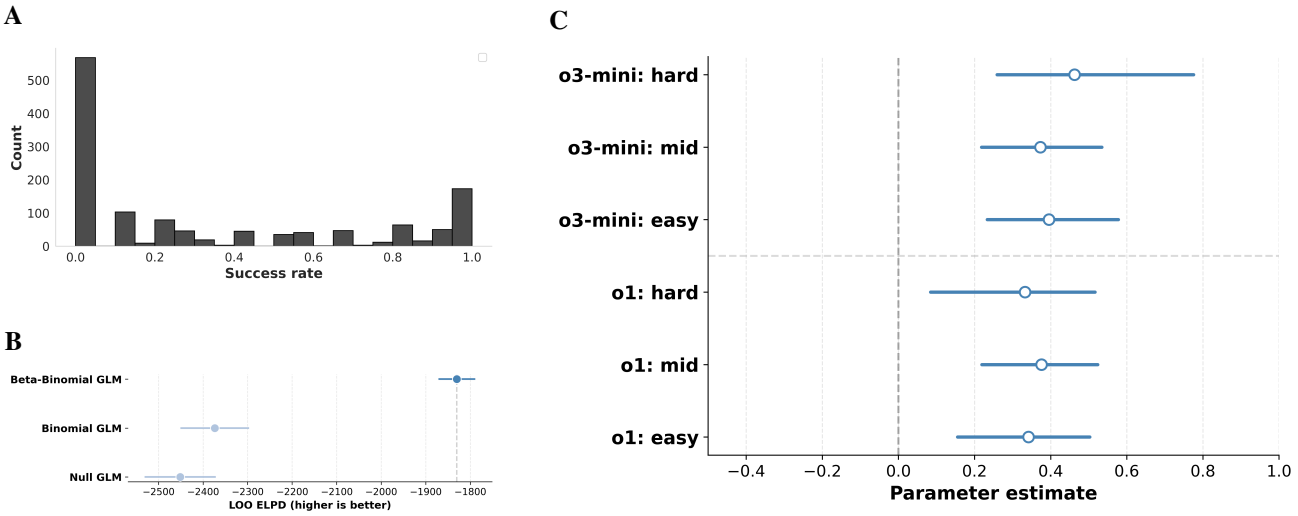

Figure 4. **GAIA agentic benchmark results. A)** Success rate distribution in GAIA. Average success rates across repeats on different tasks (across all models, difficulty levels, and reasoning efforts) are asymmetric and bimodal: models either pass or fail all attempts on most tasks. **B)** GLM comparison using LOO cross-validation. Values closer to 0 indicate better predictive accuracy. The Reasoning Beta-Binomial GLM, incorporating model-specific and difficulty-specific reasoning effort effects, provides a substantially better account of the data than the Reasoning Binomial GLM (not accounting for task-level variation in success probability) and the Null GLM (not incorporating reasoning effort effects). **C)** Forest plot for the effect of reasoning effort and task difficulty (easy, mid, hard) on performance of o1 and o3-mini from the Reasoning Beta-Binomial GLM. Posterior means (dots) and 95% HPDIs (horizontal lines) are shown on the logit scale. Positive values indicate that increased reasoning effort improves performance.

diagnostic resolution (Figure 4A).

To formally evaluate whether reasoning effort is a relevant source of variation and to assess whether 0/1 overdispersion is a factor to be considered in the data distribution, we formally compared three hierarchical GLMs using leave-one-out cross-validation (LOO; Figure 4B): (1) a *Reasoning Beta-Binomial* GLM incorporating model, difficulty, reasoning effort, and task-level random effects with Beta-Binomial likelihood (Equation (4)); (2) a *Reasoning Binomial* GLM that is structurally identical but uses a standard Binomial likelihood (Equation (5)); and (3) a *Null* GLM that omits reasoning effort entirely (Equation (6)). The formal GLM comparison yields two findings simultaneously: the Reasoning Beta-Binomial GLM substantially outperforms both alternatives, establishing that (a) reasoning effort is a meaningful predictor of performance and (b) task-level overdispersion must be accounted for (Figure 4B). This kind of inference about the data-generating process, formally comparing competing hypotheses about what drives evaluation outcomes, is not possible with t-tests or distribution-free approaches like cluster-based bootstrap methods.

Having established that the Beta-Binomial GLM provides the best account of the data, we examine its parameter estimates in detail. Figure 4C displays the posterior distributions for the linear effect of reasoning effort for each model at each difficulty level. For both o1 and o3-mini, reasoning effects are positive and reliably non-zero across all difficulty levels, all HPDIs exclude zero for easy, mid, and hard tasks. This indicates that investing in reasoning effort consistently improves performance regardless of task difficulty. However, the magnitude of the effect varies: o3-mini shows some small gains for harder task vs easy task, but not o1, suggesting that the return on reasoning investment is model-dependent even when the direction of the effect is consistent. Results for Claude-3.7-Sonnet and GPT-4.5-preview, which were tested only without reasoning, are omitted as they cannot inform reasoning effort effects. These estimates carry direct resource allocation implications. Increased reasoning effort is justified for both o1 and o3-mini across all difficulty levels, but the expected performance gain per unit of additional token cost is larger for o3-mini, particularly on hard tasks where the marginal benefit is greatest.

Such nuanced, model-specific conclusions about where to allocate expensive compute require both stable point estimates and calibrated uncertainty. In agentic evaluation settings where each additional epoch of data collection may cost substantial amounts of money and take hours, if not days to complete, researchers must often make decisions based on early, incomplete data. The key question becomes: *can I trust my current estimates enough to act on them, or will they change substantially as more data arrives?* To test this, we simulate a scenario of drawing conclusions from limited data by comparing reasoning effect estimates obtained from 3 epochs against those from the full 10 epochs collected on the benchmark.

Using the stability metrics defined in Section 3.2, we com-

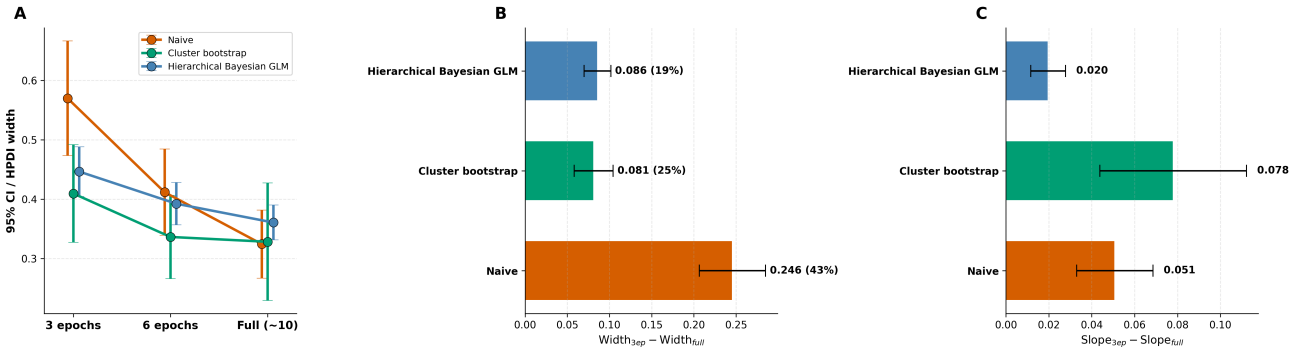

Figure 5. **Stability of reasoning effect estimates across data regimes. A)** Mean 95% CI/HPDI width ($\pm$ SEM across model–difficulty cells) for reasoning effort slopes at three data regimes (3, 6, and full $\sim$10 epochs) and uncertainty quantification approaches: Frequentist GLM naive CIs (orange), Frequentist GLM cluster bootstrap (green) and Hierarchical Bayesian GLM (blue). **B)** Width reduction from 3 epochs to full data (mean $\pm$ SEM). Larger values indicate greater confidence reduction (narrower CIs/HPDIs) in estimates with more data. **C)** Point estimate drift: change in reasoning slope mean estimate from 3 epochs to full data (mean $\pm$ SEM).

pare three approaches to uncertainty quantification for the key parameter of interest (reasoning slope). Naive CIs were extracted from a separate Frequentist Binomial GLM per model–difficulty cell (non-hierarchical), regressing task-level success counts on reasoning effort as a continuous predictor, and construct 95% confidence intervals from the asymptotic standard errors of the slope coefficient ($\hat{\beta} \pm 1.96 \cdot \text{SE}$). Cluster bootstrap resample tasks with replacement within each model–difficulty cell, re-aggregated the Binomial counts, and refitted the same GLM on each resample to obtain a bootstrap distribution of slopes (1,000 replicates); this is a distribution-free uncertainty quantification method that respects the clustering of repeated observations within tasks. Hierarchical Bayesian Beta-Binomial GLMs estimate all model–difficulty–reasoning slopes jointly with partial pooling across tasks, difficulty levels, and models via MCMC.

Figure 5 summarizes the results across three methods. The methods differ markedly in how their estimates change as data accumulates. Naive CIs exhibit the qualitatively largest change in interval width, shrinking by 43% from 3 to 10 epochs. This suggests that conclusions drawn from low amounts of data would entail substantially different uncertainty bounds than those drawn from the full dataset. Point estimates also shift meaningfully (drift: 0.051 ± 0.018). The cluster bootstrap shows more consistent interval widths across regimes (20% reduction), but its point estimates are the most sensitive to data regime of all methods (drift: 0.078 ± 0.034). Relying on cluster bootstrap slopes from 3 epochs would, on average, arrive at a substantially different estimate than one with access to the full data. Hierarchical Bayesian GLMs show interval width changes comparable to the cluster bootstrap (20% reduction), but its point estimates are the most stable across data regimes (drift: 0.019 ± 0.008), approximately $4\times$ less sensitive to data regime than the cluster bootstrap and $2.5\times$ less than naive CIs. This stability arises from partial pooling across tasks, difficulty levels, and models, which regularizes estimates even when per-cell data is scarce. In agentic evaluation settings where resource allocation decisions must be often made based on early, expensive initial runs, the sensitivity of an estimator to data regime directly affects the reliability of those decisions. We note that the standard errors on the drift estimates are non-trivial (Hierarchical Bayesian GLM: 0.019 ± 0.008; cluster bootstrap: 0.078 ± 0.034), so the precise multiplier should be interpreted with caution. The qualitative ordering, Hierarchical Bayesian GLMs most stable, naive CIs intermediate, cluster bootstrap least stable, is robust across individual model-difficulty cells, but the exact ratio will vary across settings.

We note that the stability of hierarchical Bayesian estimates could in principle reflect prior regularization rather than information sharing. However, our mildly informative priors allow substantial posterior variation, the model recovers meaningful model-specific effects, and HPDIs narrow with additional data, all consistent with data-driven rather than prior-dominated estimation. A formal prior sensitivity analysis is an important direction for future work.

In combination, the results of use case 3 illustrate two properties of hierarchical Bayesian modeling that are particularly relevant for agentic evaluation. First, formal model comparison allows researchers to adjudicate between competing hypotheses about the data-generating process, here establishing both the relevance of reasoning effort and the importance of accommodating task-level overdispersion. Conventional statistical tests do not provide a mechanism for such comparisons. Second, the stability analysis reveals that the three considered methods differ substantially in how their estimates respond to different data regimes. Among the approaches compared, hierarchical Bayesian modeling produces the most stable point estimates across

data regimes while maintaining consistent interval widths, a combination that neither naive CIs (whose widths change substantially) nor the cluster bootstrap (whose point estimates shift most) achieve. In agentic evaluation settings where data collection is expensive and resource allocation decisions depend on early estimates of performance drawn from limited amounts of data, the degree to which an estimator's conclusions change with additional data is a practically important consideration.

5. Discussion

Across three evaluation scenarios, we showed that the choice of statistical method has substantive consequences for conclusions about model capabilities. Methods not accounting for data hierarchy yield overconfident, sample size-driven inferences; the cluster bootstrap respects hierarchy but produces unstable estimates with few clusters; and hierarchical Bayesian GLMs provide the most stable estimates while uniquely supporting formal model comparison.

A distinctive advantage of hierarchical Bayesian GLMs is formal model comparison. In use case 3, this establishes that reasoning effort predicts agentic performance and that task-level overdispersion must be modeled, inferences about the data-generating process that neither t-tests nor bootstrap methods can provide.

The stability analysis further shows that partial pooling produces estimates robust to data scarcity: reasoning slopes from 3 epochs change minimally with the full dataset, whereas cluster bootstrap slopes shift substantially. As agentic evaluation costs rise (Kamradt, 2025), drawing reliable conclusions from limited data is an economic necessity.

A potential barrier to adoption of hierarchical Bayesian methods is perceived implementation complexity. Modern probabilistic programming frameworks make specifying and fitting multilevel GLMs increasingly straightforward, and several open-source tools now support streamlined workflows from evaluation log files to posterior summaries with uncertainty intervals. Hierarchical Bayesian GLMs also require specifying prior distributions, introducing researcher degrees of freedom that must be transparently reported. While we use mildly informative priors throughout, prior sensitivity analysis is an important direction for future work.

MCMC estimation is computationally more expensive than closed-form Frequentist methods, but all models here were estimated in under 5 minutes on 4 CPUs, with further speedups available via GPU acceleration. The analyses presented focus on binary (pass/fail) outcomes; extensions to ordinal or continuous outcome variables are straightforward within the GLM framework (Dubois et al., 2025) but not demonstrated here. The models presented here can also be fit in a Frequentist mixed-effects framework, which shares the partial-pooling property of hierarchical Bayesian GLMs and would likely yield similar point estimates in data-rich settings. However, the advantages of our approach that we emphasize, full posterior distributions for uncertainty quantification, formal model comparison via LOO-CV, and the potential for sequential updating, are specific to the Bayesian framework. A direct empirical comparison between the two GLM approaches is a valuable direction for future work. Lastly, our stability analysis compares 3 versus 10 epochs on a single benchmark; a broader empirical evaluation across multiple agentic benchmarks would strengthen claims about generalizability.

Future directions. The Bayesian framework naturally supports *sequential updating*, where posteriors from early runs serve as priors for subsequent data, enabling adaptive evaluation protocols. Consideration of hierarchical data structure can also be leveraged to estimate *cross-task correlations* for predictive evaluations, and jointly model inexpensive QA benchmarks with costly agentic evaluations to borrow strength across modalities, potentially reducing evaluation burden.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks, 2025. URL https://arxiv.org/abs/2511.04703.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the clt in llm evals with fewer than a

few hundred datapoints. 3 2025. URL http://arxiv.org/abs/2503.01747.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3): 414–427, 2008.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. 5 2019. URL http://arxiv.org/abs/1905.10044.

Harry Coppock, Lennart Luettgau, Magda Dubois, Christopher Summerfield, and Cozmin Ududec. Hibayes: A python package for analysing data from inspect logs using statistical modeling techniques. https://github.com/UKGovernmentBEIS/hibayes, 2025.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Magda Dubois, Harry Coppock, Mario Giulianelli, Timo Flesch, Lennart Luettgau, and Cozmin Ududec. Skewed score: A statistical framework to assess autograders, 2025. URL https://arxiv.org/abs/2507.03772.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. 3rd edition, 2013.

Tianyu Han, Daniel Watson, Manish Raghavan, Soumya Patil, Subhro Ramos, Larry Tang, Jiani Choi, Jonghyun Kim, Inwoo Kim, Zhengbao Zheng, Jianfeng Yang, Hanjun Pan, Arun Kamath, Kenton Lee, Kush Varshney, Percy Liang, Kathleen Lee, Karthik Narasimhan, Peter Clark, and Luke Zettlemoyer. Swe-bench: Can language models resolve real-world github issues? https://github.com/SWE-bench, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. 9 2020. URL http://arxiv.org/abs/2009.03300.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. 11 2011. URL http://arxiv.org/abs/1111.4246.

Greg Kamradt. Analyzing openai's o1-preview in comparison with claude 3 opus and gpt-4 on arc agi evaluation, 2025. URL https://arcprize.org/blog/analyzing-o3-with-arc-agi.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. 4 2017. URL http://arxiv.org/abs/1704.04683.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Wen tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. 11 2022. URL http://arxiv.org/abs/2211.11501.

Mary Llewellyn, Annie Gray, Josh Collyer, and Michael Harries. Towards reliable and practical LLM security evaluations via Bayesian modelling, 2025. URL https://arxiv.org/abs/2510.05709.

Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior, 2025. URL https://arxiv.org/abs/2511.10661.

Richard McElreath. *Statistical Rethinking*. CRC Press, 2nd edition, 2020. doi: 10.1080/09332480.2017.1302722.

Richard McElreath. Statistical rethinking 2023. https://github.com/rmcelreath/stat_rethinking_2023, 2023.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. 11 2023. URL http://arxiv.org/abs/2311.12983.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. 11 2024. URL http://arxiv.org/abs/2411.00640.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. 12 2019. URL http://arxiv.org/abs/1912.11554.

David Rein, Joel Becker, Amy Deng, Seraphina Nix, Chris Canal, Daniel O'connell, Pip Arnott, Ryan Bloom, Thomas Broadley, Katharyn Garcia, Brian Goodrich, Max Hasin, Sami Jawhar, Megan Kinniment, Thomas Kwa, Aron Lajko, Nate Rush, Lucas Jun, Koba Sato, Sydney Von Arx, Ben West, Lawrence Chan, and Elizabeth Barnes. Hcast: Human-calibrated autonomy software tasks. 3 2025. URL https://arxiv.org/abs/2503.17354.

Yaniv Romano, Evan Patterson, and Emmanuel J Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. 10 2022. URL http://arxiv.org/abs/2210.03057.

UKGovernmentBEIS. inspect_ai: AI-assisted inspection policy development. https://github.com/UKGovernmentBEIS/inspect_ai, 2024.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. 6 2023. URL http://arxiv.org/abs/2306.14898.

Zhaojian Yu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. Humaneval pro and mbpp pro: Evaluating large language models on self-invoking code generation. 12 2024. URL http://arxiv.org/abs/2412.21199.

6. Appendix

A. Model specifications

All hierarchical Bayesian GLMs use non-centered parameterizations to improve MCMC sampling efficiency. Parameters are estimated using NUTS (Hoffman and Gelman, 2011) via NumPyro (Phan et al., 2019) with 2,000 warmup and 2,000 inference samples across 4 chains.

A.1. Use case 1: Single model, multiple domains

Data is aggregated to benchmark level: for each benchmark j within domain d, we observe Correct_j successes out of n_j trials.

$$\begin{aligned}
\text{Correct}_j &\sim \text{Binomial}(n_j, p_j) \\
\text{logit}(p_j) &= \text{domain}_{d[j]} \\
\text{domain}_d &\sim \text{Normal}(\bar{\mu}, \sigma_{\text{domain}}) \\
\bar{\mu} &= \mu_{\text{overall}} + \sigma_{\text{overall}} \cdot z_{\text{overall}} \\
\mu_{\text{overall}} &\sim \text{Normal}(0, 1) \\
\sigma_{\text{overall}} &\sim \text{HalfNormal}(0.5) \\
z_{\text{overall}} &\sim \text{Normal}(0, 1) \\
\sigma_{\text{domain}} &\sim \text{HalfNormal}(0.1)
\end{aligned} \tag{2}$$

Correct_j denotes the number of correct answers for benchmark j, modeled as a Binomial draw with success probability p_j among n_j trials. Domain-specific effects are drawn from a Normal distribution whose mean $\bar{\mu}$ is itself drawn from a reparameterized hyper-distribution governed by μ_{overall}, σ_{overall}, and z_{overall}, using mildly informative priors. The overall performance estimate is obtained from the hyper-mean $\bar{\mu}$, which regularizes across domains via partial pooling.

A.2. Use case 2: Multiple models, multiple domains

Data is aggregated to the (model $\times$ domain) level. The hierarchy is: overall $\rightarrow$ model $\rightarrow$ (model $\times$ domain).

$$\begin{aligned}
\text{Correct}_{i,j} &\sim \text{Binomial}(n_{i,j},\ p_{i,j}) \\
\text{logit}(p_{i,j}) &= \text{domain}_{i,j} \\
\mu_{\text{overall}} &\sim \text{Normal}(0, 1.5) \\
\sigma_{\text{model}} &\sim \text{HalfNormal}(0.1) \\
z_{\text{model},i} &\sim \text{Normal}(0, 1) \\
\text{model}_i &= \mu_{\text{overall}} + \sigma_{\text{model}} \cdot z_{\text{model},i} \\
\sigma_{\text{domain}} &\sim \text{HalfNormal}(1.0) \\
z_{\text{domain},i,j} &\sim \text{Normal}(0, 1) \\
\text{domain}_{i,j} &= \text{model}_i + \sigma_{\text{domain}} \cdot z_{\text{domain},i,j}
\end{aligned} \tag{3}$$

$\text{Correct}_{i,j}$ denotes the number of correct answers for the i-th model in the j-th domain, modeled as a Binomial draw with success probability $p_{i,j}$ among $n_{i,j}$ trials. Model-level effects are drawn from a distribution centered on the overall mean. Domain-level effects are drawn from a distribution centered on the respective model effect, enabling partial pooling across domains within each model and across models at the overall level.

A.3. Use case 3: Agentic evaluation with reasoning effort

Data is aggregated to the (model $\times$ task $\times$ reasoning level $\times$ difficulty) cell level. We specify three competing GLMs.

A.3.1. Reasoning Beta-Binomial GLM

$$
\begin{aligned}
\text{Correct}_{i,j,k,d} &\sim \text{BetaBinomial}(n_{i,j,k,d},\ \alpha_{i,j,k,d},\ \beta_{i,j,k,d}) \\
\text{logit}(p_{i,j,k,d}) &= \text{model}_i + \text{difficulty}_d + \text{task}_j \\
&\quad + \gamma_{i,d} \cdot \text{reasoning}_k \\
\alpha_{i,j,k,d} &= p_{i,j,k,d} \cdot \varphi \\
\beta_{i,j,k,d} &= (1 - p_{i,j,k,d}) \cdot \varphi \\
\varphi &\sim \text{Gamma}(2, 0.5) \\
\mu_{\text{overall}} &\sim \text{Normal}(0, 1.5) \\
\sigma_{\text{model}} &\sim \text{HalfNormal}(0.5) \\
z_{\text{model},i} &\sim \text{Normal}(0, 1) \\
\text{model}_i &= \mu_{\text{overall}} + \sigma_{\text{model}} \cdot z_{\text{model},i} \\
\sigma_{\text{difficulty}} &\sim \text{HalfNormal}(0.5) \\
z_{\text{difficulty},d} &\sim \text{Normal}(0, 1) \\
\text{difficulty}_d &= \sigma_{\text{difficulty}} \cdot z_{\text{difficulty},d} \\
\sigma_{\text{task}} &\sim \text{HalfNormal}(0.5) \\
z_{\text{task},j} &\sim \text{Normal}(0, 1) \\
\text{task}_j &= \sigma_{\text{task}} \cdot z_{\text{task},j} \\
\mu_\gamma &\sim \text{Normal}(0, 0.5) \\
\sigma_\gamma &\sim \text{HalfNormal}(0.5) \\
z_{\gamma,i,d} &\sim \text{Normal}(0, 1) \\
\gamma_{i,d} &= \mu_\gamma + \sigma_\gamma \cdot z_{\gamma,i,d}
\end{aligned}
\tag{4}
$$

$\text{Correct}_{i,j,k,d}$ denotes the number of correct answers for the i-th model, j-th task, k-th reasoning level, and d-th difficulty level. The Beta-Binomial likelihood allows each observation's success probability to vary around $p_{i,j,k,d}$ according to a Beta distribution controlled by the concentration parameter φ, accommodating the overdispersion (bimodal success rates) typical of agentic evaluations. The reasoning effect $\gamma_{i,d}$ is a linear slope on the ordinal reasoning level (0–3), estimated separately for each model–difficulty combination with partial pooling via the shared hyper-parameters μ_γ and σ_γ.

A.3.2. Reasoning Binomial GLM

$$
\begin{aligned}
\text{Correct}_{i,j,k,d} &\sim \text{Binomial}(n_{i,j,k,d},\ p_{i,j,k,d}) \\
\text{logit}(p_{i,j,k,d}) &= \text{model}_i + \text{difficulty}_d + \text{task}_j \\
&\quad + \gamma_{i,d} \cdot \text{reasoning}_k
\end{aligned}
\tag{5}
$$

Structurally identical to the Beta-Binomial GLM above, but using a standard Binomial likelihood. All hierarchical priors on model, difficulty, task, and reasoning effects are the same as in Equation (4). This model does not account for task-level overdispersion.

A.3.3. Null GLM

$$
\begin{aligned}
\text{Correct}_{i,j,d} &\sim \text{Binomial}(n_{i,j,d},\ p_{i,j,d}) \\
\text{logit}(p_{i,j,d}) &= \text{model}_i + \text{difficulty}_d + \text{task}_j
\end{aligned}
\tag{6}
$$

Same hierarchical structure as the Reasoning Binomial GLM but without the reasoning slope $\gamma_{i,d}$. All hierarchical priors on

Table 2. **MCMC convergence diagnostics for all fitted hierarchical Bayesian GLMs.** For each model, we report the maximum split-$\hat{R}$ across all parameters, the minimum bulk and tail effective sample sizes (ESS), and the number of divergent transitions (out of 8,000 total post-warmup samples across 4 chains). All models show adequate convergence.

	Model	Data	$\hat{R}$	Min bulk ESS	Min tail ESS	Divergences
Use case 1	Binomial GLM	Full	1.00	1,807	2,039	0
Use case 2	Binomial GLM	Full	1.00	1,471	2,119	0
Use case 3	Reasoning Binomial	3ep	1.00	1,539	2,040	0
	Reasoning Beta-Binomial	3ep	1.00	2,339	2,339	0
	Null Binomial	3ep	1.00	1,441	1,691	0
	Reasoning Binomial	6ep	1.00	1,046	1,878	0
	Reasoning Beta-Binomial	6ep	1.00	1,824	2,203	1
	Null Binomial	6ep	1.00	997	1,042	0
	Reasoning Binomial	Full	1.00	514	1,204	0
	Reasoning Beta-Binomial	Full	1.01	1,481	1,641	1
	Null Binomial	Full	1.01	567	1,043	0

model, difficulty, and task effects are the same. This model encodes the null hypothesis that reasoning effort does not affect performance.

A.4. Frequentist comparison methods for use case 3

For the stability analysis in use case 3, we compare hierarchical Bayesian modeling against two frequentist alternatives that estimate reasoning effect slopes per model–difficulty cell.

Naive CI. For each (model $\times$ difficulty) cell, data is aggregated to the (reasoning level $\times$ task) level, yielding Binomial counts (successes out of total attempts). An aggregated Binomial GLM is fit via maximum likelihood:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \text{reasoning} \tag{7}$$

where reasoning $\in \{0, 1, 2, 3\}$ is the ordinal reasoning level. The naive 95% CI is $\hat{\beta}_1 \pm 1.96 \cdot \text{SE}(\hat{\beta}_1)$, using the asymptotic standard error from the Fisher information matrix.

Cluster bootstrap. Within each (model $\times$ difficulty) cell, tasks are resampled with replacement (preserving all reasoning levels within each resampled task). For each of 1,000 replicates, the data is re-aggregated and the same Binomial GLM is refit. The 95% interval is taken from the 2.5th and 97.5th percentiles of the bootstrap slope distribution. This is a distribution-free method that respects the clustering of repeated observations within tasks.

A.5. Convergence diagnostics

For all hierarchical Bayesian GLMs reported in the main text, we verified MCMC convergence using standard diagnostics. Table 2 summarizes the maximum split-$\hat{R}$, minimum bulk and tail effective sample sizes (ESS), and number of divergent transitions across all parameters for each fitted model. All models satisfy standard convergence criteria: $\hat{R} \leq 1.01$ for all parameters, minimum bulk ESS ≥ 500, and near-zero divergent transitions.

A.6. Use case 2: Shrinkage

The conclusion of no evidence for model differences warrants scrutiny: strong partial pooling across few domains could shrink genuinely different model effects toward the grand mean, producing apparent equivalence as an artifact of the hierarchical structure. To assess this, we examine the posterior for σ_{model}, which controls the degree of between-model variation. The posterior mean of σ_{model} is 0.078 (89% ETI $[0.007, 0.19]$; reported as 89% ETI from default summary output, consistent conclusions hold under 95% HPDIs). This small value indicates that the data support very limited between-model variation on the logit scale: even at the upper end of the credible interval, this corresponds to roughly ± 4 percentage points around the overall mean. The equivalence finding thus reflects the data's evidence that between-model variation is small relative to between-domain variation (σ_{domain} posterior mean: 0.90, 89% ETI $[0.64, 1.3]$), rather than an artifact of excessive

shrinkage. Domain-level differences between models do exist and can be extracted from the fitted model, for instance, Sonnet-4 achieves 0.95 on reasoning versus GPT-4o's 0.83, but these differences are absorbed by the domain-level variance component rather than appearing at the model level.