ENTERPRISE'S LAST EXAM: A BENCHMARK FOR ORGANIZATIONAL JUDGMENT UNDER FRAGMENTED ENTERPRISE CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier language models perform strongly on benchmarks of coding, mathematics, knowledge, and general reasoning. However, enterprise deployments often require a different capability: making decisions whose correctness depends on organization-specific entities, historical precedent, conflicting systems of record, changing policy versions, distributed or informal authority, and temporal context. We introduce Enterprise's Last Exam (ELE), a benchmark for this capability, which we term organizational judgment. ELE complements benchmarks of generic knowledge, artifact production, and software interaction by directly measuring whether models identify and correctly apply the organization-specific evidence that governs a decision. ELE comprises 630 practitioner-authored scenarios across three complementary model-blind evaluation sets: a Core set ($n = 384$), a difficulty-focused Challenge set ($n = 156$), and 90 counterfactual scenarios forming 45 paired cases in which a single decision-critical fact changes. Scenarios span six dimensions: entity resolution under ambiguity, precedent-based exception handling, cross-system synthesis, policy-version reasoning, approval-chain reconstruction, and temporal decision consistency. We evaluate seven proprietary and open-weight models under a controlled direct-prompt condition at temperature zero. On the primary Core set, accuracy ranges from 77.4% to 97.4%; no model saturates the benchmark. Performance varies sharply by reasoning type: models are strongest on entity resolution, approval-chain reconstruction, and temporal consistency, but substantially weaker when they must apply precedent or reconcile conflicting evidence across systems; cross-system synthesis ranges from 67% to 90%. Challenge accuracy ranges from 73.7% to 87.8%. On the 45 counterfactual pairs, the fraction for which models answer both variants correctly ranges from 62% to 89%, indicating substantial variation in robustness when a single governing organizational fact changes.

1 INTRODUCTION

Large language models are increasingly evaluated on tasks that resemble economically valuable work rather than only academic examinations. Recent benchmarks assess web-based knowledge work, CRM workflows, tool use, professional deliverables, and long-horizon agentic tasks. These evaluations have substantially improved our understanding of whether models can produce useful work products or complete workflows in realistic software environments. Yet an important class of enterprise decisions remains comparatively under-measured.

Many consequential workplace decisions are difficult not because they require obscure world knowledge or long mathematical derivations, but because the correct action is specific to an organization. A customer appearing under two slightly different account names may or may not be the same operative entity. An exception approved last quarter may establish a precedent, or may have been a one-time accommodation. A policy may have changed between commitment and fulfillment. The person with the highest formal title may not possess the relevant approval authority. A CRM record, billing system, contract, email thread, and policy document may each be individually accurate yet appear contradictory until their timestamps and scopes are reconciled.

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 refer to the capability required to resolve such cases as *organizational judgment*: selecting an action by identifying, reconciling, and applying the organization-specific evidence that governs a decision. Organizational judgment differs from generic domain expertise. A model may know how discounts, procurement approvals, support escalations, or renewal policies generally work and still make the wrong decision for a particular firm. It also differs from task execution. An agent may be perfectly capable of operating enterprise software while acting on an incorrect interpretation of who the customer is, which policy applies, or who has authority.

This distinction matters as frontier models improve. If evaluation begins only after the relevant organizational context has already been cleanly specified, benchmark performance may overestimate readiness for settings where determining the governing context is itself the core problem. Conversely, a benchmark assembled by retaining only cases that current models fail would underestimate average capability. A useful evaluation should therefore isolate the construct, avoid performance-conditioned selection, and test whether models remain sensitive to decision-critical organizational evidence.

ELE operationalizes this problem using 630 scenarios across six reasoning dimensions and three model-blind evaluation sets. The primary Core set contains 384 broadly balanced scenarios; the Challenge set contains 156 scenarios curated to emphasize harder organizational reasoning; and the Counterfactual set contains 90 scenarios forming 45 paired cases in which changing one governing fact changes the correct action. All three sets were constructed without knowledge of how the seven evaluated models would perform.

Our empirical results show that organizational judgment is not a single scalar capability. Core accuracy spans 77.4% to 97.4%. Across models, dimensions like approval-chain reconstruction, entity resolution, and temporal consistency average above 91%, whereas policy-version reasoning, precedent-based exception handling, and cross-system synthesis are substantially harder. These three form a cluster of the hardest dimensions on average, and per-category accuracy varies widely across models, with policy-version reasoning spanning 63% to 93%. These differences are larger than several of the gaps between model-level aggregate scores.

We make four contributions. First, we define organizational judgment as a distinct evaluation construct and provide a six-dimension taxonomy. Second, we introduce a 630-scenario benchmark with model-blind Core, Challenge, and Counterfactual sets. Third, we evaluate seven proprietary and open-weight models under one controlled direct-prompt condition. Fourth, we provide category-level and counterfactual analyses that reveal which forms of organization-specific reasoning remain most fragile.

2 Related Work

Academic capability benchmarks. MMLU established a broad measure of language-model knowledge and reasoning across academic and professional subjects (Hendrycks et al., 2021). Humanity's Last Exam (HLE) deliberately targets the frontier of expert-level academic difficulty using expert-authored, closed-ended questions across many disciplines (Center for AI Safety et al., 2026). These benchmarks measure broad knowledge and difficult reasoning; ELE instead focuses on decisions whose correctness depends on local organizational evidence.

Enterprise software and agent benchmarks. WorkArena evaluates browser agents on common knowledge-work tasks in ServiceNow, while WorkArena++ expands to hundreds of realistic compositional workflows requiring planning, retrieval, contextual understanding, and reasoning (Drouin et al., 2024; Boisvert et al., 2024). CRMArena evaluates agents on realistic CRM tasks grounded in interconnected business objects and professional workflows (Huang et al., 2025). These benchmarks ask whether agents can operate software and complete business tasks. ELE is complementary: its primary unit is the organization-specific decision that should guide action.

Tool use and policy following. τ-bench evaluates language agents interacting with simulated users, APIs, and domain-specific policies, with end-state correctness and repeated-trial reliability as central measures (Yao et al., 2025). ELE targets a related but earlier failure point: identifying which policy, precedent, entity, authority path, or time slice governs the case before execution begins.

Professional knowledge work. GDPval evaluates economically valuable knowledge-work deliverables across 44 occupations and nine major U.S. economic sectors (Patwardhan et al., 2026). APEX-Agents evaluates 480 long-horizon, cross-application tasks authored by investment bankers, management consultants, and corporate lawyers (Vidgen et al., 2026). ELE narrows the focus from end-to-end professional output to whether the model infers the correct organization-specific action from fragmented or relational evidence.

Latent organizational reasoning. ENTLORE is closest in motivation. It studies enterprise question answering when organizational relations remain implicit across heterogeneous records and introduces a graph-grounded benchmark with 2,341 documents and 907 questions spanning explicit lookup, cross-source composition, and latent organizational reasoning (Zheng et al., 2026). ENTLORE emphasizes recovering latent relations from a document corpus; ELE emphasizes applying organization-specific relations (precedent, policy version, entity identity, authority, and temporal commitments) to choose an action.

Contamination-resistant evaluation. Public benchmark reuse creates contamination and selection risks. LiveBench mitigates contamination through refreshed, objectively scored questions (White et al., 2025), while Benchmark Inflation demonstrates that public benchmark scores can exceed performance on matched retro-holdouts (Haimes et al., 2024). ELE addresses a different but related threat: performance-conditioned item selection. All reported ELE sets were constructed without observing target-model performance.

Taken together, prior work shows rapid movement from static academic tests toward realistic knowledge work, enterprise software, agentic workflows, and latent organizational structure. ELE does not claim that organizational judgment is absent from these benchmarks; rather, it measures the construct directly and decomposes it into failure modes that aggregate task-completion metrics can obscure.

3 Organizational Judgment

3.1 Definition and scope

We define organizational judgment as the ability to select a decision or action by correctly identifying, reconciling, and applying organization-specific evidence whose relevance depends on relationships among entities, systems, policies, precedents, authority structures, and time. A task qualifies when generic domain norms are insufficient but the supplied organizational evidence determines a defensible action.

This definition separates organizational judgment from neighboring capabilities. Pure retrieval asks whether a model can locate an explicitly stated fact. Generic business reasoning asks what a typical organization should do. Tool-use benchmarks ask whether an agent can execute actions in software. Long-horizon planning asks whether a system can decompose and complete extended tasks. Organizational judgment may require any of these capabilities, but the distinguishing feature is that the correct decision is conditional on local organizational context.

3.2 Six reasoning dimensions

Table 1: ELE dimensions and governing relations.

ELE dimension	Governing relation	Typical failure
Entity resolution under ambiguity	Identity, hierarchy, legal or commercial equivalence	Conflating similar entities; failing to join aliases or parent-child records
Precedent-based exception handling	Scope and authority of prior exceptions	Treating one-off accommodation as reusable rule; ignoring controlling precedent
Cross-system synthesis	Source authority and reconciliation across records	Choosing one source naively; missing scope or timestamp differences
Policy-version reasoning	Effective date, grandfathering, event-time applicability	Applying the wrong policy version to the operative event
Approval-chain reconstruction	Delegation, thresholds, business-unit authority	Assuming formal seniority equals operative approval rights
Temporal decision consistency	Commitments, knowledge states, supersession over time	Ignoring what was known or promised at the relevant time

3.3 COUNTERFACTUAL SENSITIVITY

Ordinary accuracy cannot distinguish genuine use of governing evidence from shortcut behavior. A model may answer correctly because a phrase, option pattern, or generic norm correlates with the gold action. ELE therefore includes paired counterfactuals. Each pair changes one decision-critical fact while holding the rest of the scenario as constant as possible, and the changed fact is defined so that the correct organizational action must change.

Examples of such interventions include moving a policy effective date across a transaction date, changing an account mapping, modifying delegated approval authority, changing which system is authoritative, or replacing a reusable precedent with an explicitly one-off exception. The present study reports both per-scenario accuracy and pair-level success on the 45 counterfactual pairs.

4 BENCHMARK CONSTRUCTION

4.1 SCENARIO PROVENANCE AND COMPOSITION

ELE scenarios are derived from recurring patterns in enterprise work. They were drafted with practitioner-informed templates, then reviewed and validated by the authors, and rewritten to be self-contained and privacy-preserving. Authors abstract away confidential names, identifiers, proprietary data, and unnecessary employer-specific details while retaining the relational structure that makes the decision nontrivial. Each scenario contains a decision prompt, the evidence required to answer it, a gold decision, and a rationale identifying the facts that govern that decision.

The final benchmark contains 630 scenarios spanning the six reasoning dimensions above and eight organizational domains: finance and revenue operations, compliance and legal, sales and deal desk, HR and people operations, procurement and vendor management, customer success and support, engineering and DevOps, and a small set of cross-functional cases. The benchmark samples recurring organizational decision patterns rather than reproducing individual corporate incidents verbatim.

4.2 THREE COMPLEMENTARY MODEL-BLIND SETS

Table 2: ELE evaluation sets.

Set	Scenarios	Role	Construction
Core	384	Primary capability estimate	Broad, balanced coverage across all six reasoning dimensions
Challenge	156	Difficulty-focused diagnostic	Cases curated to emphasize harder organizational reasoning, assigned without screening on evaluated-model failures
Counterfactual	90 (45 pairs)	Evidence sensitivity	One decision-critical fact changes and the correct action must change

All three sets are model-blind with respect to the reported evaluation: their scenarios were created and assigned to sets without knowledge of how the seven target models would perform. This distinction matters for the Challenge set. Its cases were curated to emphasize harder organizational reasoning, but the set was not selected after observing target-model errors, so it is not tuned in response to any target model's errors. Core provides the headline capability estimate. Challenge is reported as a separate diagnostic, and its accuracy is not assumed to fall below Core; Section 6.2 shows that it does not do so uniformly.

4.3 SCORING

ELE prioritizes decision correctness. Responses are first scored by exact match when the required answer can be deterministically extracted. Every non-exact response is graded by a fixed LLM-as-judge pipeline that scores the response against the held-out correct answer and rationale on a scale from 0 to 1, with a strict correctness threshold of 0.9; empty or non-generated responses are scored incorrect without a judge call. A response that discusses relevant concepts but recommends the wrong organizational action receives no partial credit. This keeps the primary metric aligned with operational

decision quality rather than semantic similarity. The judge is a separate fixed model (claude-opus-4-6) applied identically to all seven evaluated systems. In the reported run, exact match resolved 73.5% of responses and the remaining 26.5% followed the judge path; the per-model non-exact share ranged from 22.7% to 32.4%. Supplementary materials provide the artifacts needed to reproduce the reported aggregate results and audit the evaluation and scoring pipeline; access to held-out gold answers is described in the Reproducibility Statement.

The reported run completed without execution or scoring errors. Because all seven models were evaluated on all 630 scenarios, the design contains 4410 model-scenario evaluations. We report the benchmark's supplied accuracy values and 95% Wilson confidence intervals for the overall and Core aggregates, with selected category-level intervals in Section 6.3.

4.4 Counterfactual construction

The Counterfactual set contains 45 base/counterfactual pairs (90 scenarios). Within each pair, a single decision-critical organizational fact is changed so that the correct action changes. The paired design targets causal relevance rather than generic difficulty: a model should not merely answer each scenario plausibly, but should condition its decision on the fact that controls the organizational outcome.

5 Evaluation Protocol

5.1 Models and prompting

We evaluate a fixed snapshot of seven proprietary and open-weight systems: gpt-5.6-terra, qwen3-235b, llama4-maverick, claude-opus-5, deepseek-v3.2, gemma-3-27b, and nova-micro. Every model is run under a single controlled direct-prompt condition with temperature 0. The prompt provides the complete scenario and asks the model to select the required organizational action in the specified answer format. No category label or gold rationale is supplied. The roster is treated as an evaluation snapshot rather than a claim to include the newest release in every model family.

The single-condition design intentionally favors comparability over prompt optimization. It measures how the systems behave under one common elicitation regime rather than the best score obtainable through model-specific scaffolding, repeated sampling, or agentic tool use. Full evaluation configuration is reported in Table A1.

5.2 Metrics and reporting

Core accuracy is the primary capability estimate. Challenge and Counterfactual accuracy are reported separately because they serve different diagnostic purposes. We additionally report an overall accuracy across all 630 scenarios, category-level accuracy across the six organizational-judgment dimensions, and 95% Wilson confidence intervals for the overall and Core aggregates. Core-set confidence intervals are reported in Table A2. Because every model answers the same scenarios, model comparisons use paired data, and we report a McNemar test for the top two systems. For the Counterfactual set we report per-scenario accuracy together with pair-level success, the fraction of the 45 pairs answered correctly on both members.

We report counterfactual results at two levels. Per-scenario accuracy measures performance under decision-critical factual perturbations. Pair-level success, the fraction of pairs answered correctly on both members, tests whether a model holds up across the base and counterfactual variants of the same case. We report both and do not conflate them.

6 Results

6.1 Overall performance

No evaluated model saturates ELE. On the primary Core set, accuracy ranges from 77.4% to 97.4%. Across all 630 scenarios, overall accuracy ranges from 76.4% to 94.1%. claude-opus-5 has the highest aggregate accuracy at 94.1% (95% CI [92, 96]), followed by gpt-5.6-terra at 91.5% and qwen3-235b

at 86.0%. The 17.7-point gap between the highest and lowest overall scores indicates substantial variation across current systems, while the remaining error rate of the strongest model leaves room for meaningful progress. A paired McNemar test indicates the top-ranked claude-opus-5 is ahead of the second-ranked gpt-5.6-terra, though by a smaller margin ($\chi^2 = 4.50$, 1 df, $p \approx 0.03$, over 630 jointly scored items, with 33 discordant items favoring claude-opus-5 and 17 favoring gpt-5.6-terra).

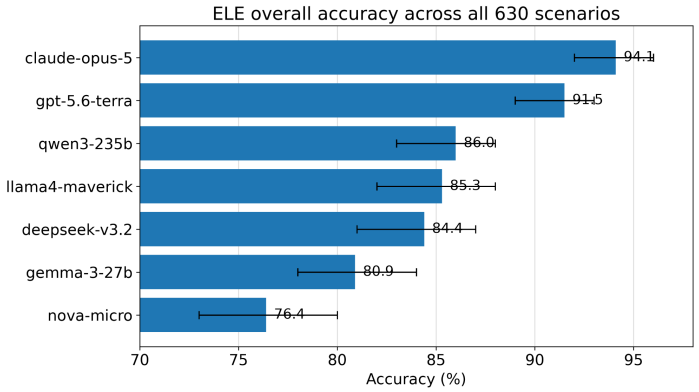

Figure 1: Overall accuracy across all 630 ELE scenarios. Error bars show the supplied 95% Wilson confidence intervals.

Table 3: Overall and split-level accuracy (%).

Model	Overall (630)	95% CI	Core (384)	Challenge (156)	Counterfactual (90)
claude-opus-5	94.1	[92, 96]	97.4	85.9	94.4
gpt-5.6-terra	91.5	[89, 93]	92.9	86.5	94.4
qwen3-235b	86.0	[83, 88]	86.6	84.0	86.7
llama4-maverick	85.3	[82, 88]	85.0	87.8	82.2
deepseek-v3.2	84.4	[81, 87]	85.8	82.1	82.2
gemma-3-27b	80.9	[78, 84]	82.9	77.6	77.8
nova-micro	76.4	[73, 80]	77.4	73.7	76.7

6.2 Core, Challenge, and Counterfactual profiles

Challenge is not simply a uniformly harder copy of Core. Relative to Core, claude-opus-5 shows the largest drop, from 97.4% to 85.9% (-11.5 points), even though it is the strongest model on Core. gpt-5.6-terra drops 6.4 points, gemma-3-27b drops 5.3, deepseek-v3.2 and nova-micro each drop 3.7, and qwen3-235b drops 2.6, while llama4-maverick improves by 2.8. Llama4-maverick records the highest Challenge score at 87.8%, even though it ranks fourth overall. The model with the best Core accuracy shows the largest drop on Challenge, reinforcing that the difficulty-focused set probes a different capability profile.

This pattern suggests that the three splits expose *different capability profiles* rather than forming a single monotonic difficulty ladder. Core remains the appropriate headline estimate because of its broad coverage, while Challenge is useful for revealing which models are relatively robust on the harder organizational decisions it emphasizes.

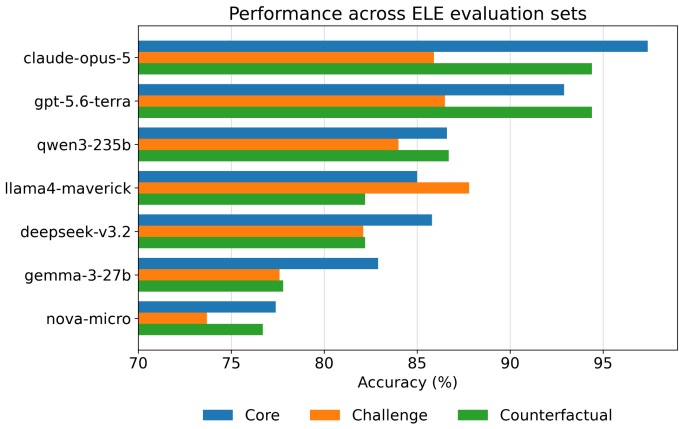

Figure 2: Accuracy on the Core, Challenge, and Counterfactual evaluation sets.

6.3 Organizational judgment is strongly dimension-dependent

Accuracy varies substantially across reasoning dimensions. Averaged across the seven models, approval-chain reconstruction is highest at 94.4%, followed by temporal consistency at 92.3% and entity resolution at 92.1%. Performance is lower on policy-version reasoning (79.1%), cross-system synthesis (77.4%), and precedent-based exception handling (76.6%). The mean difference between the highest- and lowest-scoring dimensions is therefore about 18 percentage points. Full model-by-dimension results are reported in Table A3.

The three lowest-scoring dimensions are clustered: precedent-based exception handling (76.6%), cross-system synthesis (77.4%), and policy-version reasoning (79.1%). Policy-version reasoning shows the widest cross-model spread, from 63% to 93%; cross-system synthesis ranges from 67% to 90% and precedent-based exception handling from 66% to 92%. By contrast, approval-chain reconstruction, entity resolution, and temporal consistency each span only about nine points across the field. Per-category sample sizes are 106 scenarios for entity resolution, approval-chain reconstruction, and temporal consistency, and 104 for cross-system synthesis, policy-version reasoning, and precedent-based exception handling. At these sizes a single-category 95% Wilson interval spans roughly 16 to 20 points; for example, gpt-5.6-terra's 90% on cross-system synthesis has a 95% CI of [83, 95], while nova-micro's 67% has [58, 76].

Category profiles also differ substantially by model. nova-micro, for example, scores 89% on approval-chain reconstruction but 63% on policy-version reasoning and 66% on precedent-based exception handling. Llama4-maverick and deepseek-v3.2 each reach 97% on approval-chain reconstruction despite lower overall scores than the top two models. Aggregate accuracy therefore does not fully characterize performance on individual forms of organizational judgment.

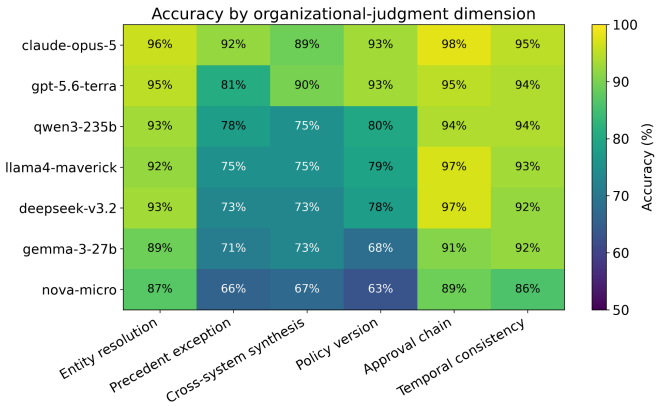

Figure 3: Category-level accuracy for the seven evaluated models. Darker cells indicate lower accuracy.

6.4 Counterfactual performance

On the 90 counterfactual scenarios, per-scenario accuracy ranges from 76.7% to 94.4%. gpt-5.6-terra and claude-opus-5 are highest at 94.4%, qwen3-235b reaches 86.7%, llama4-maverick and deepseek-v3.2 each score 82.2%, gemma-3-27b scores 77.8%, and nova-micro scores 76.7%. These results show that performance remains uneven when one governing organizational fact is changed. Because we do not yet report required-flip direction, we treat the per-scenario and pair-level numbers as evidence about robustness under controlled factual perturbations rather than as a complete causal-update score. At the pair level (Table B1), gpt-5.6-terra and claude-opus-5 answer both members correctly on 40 of 45 pairs (89%), followed by qwen3-235b (34, 76%), llama4-maverick (33, 73%), deepseek-v3.2 (31, 69%), gemma-3-27b (30, 67%), and nova-micro (28, 62%). The number of pairs on which a model answers exactly one member correctly, a direct sign of brittleness to the changed fact, ranges from 5 (gpt-5.6-terra and claude-opus-5) to 13 (nova-micro).

7 Discussion

Performance varies markedly across ELE dimensions. Approval-chain reconstruction, entity resolution, and temporal consistency are the three highest-scoring categories, whereas policy-version reasoning, precedent-based exception handling, and cross-system synthesis score lower. One possible explanation is that the higher-scoring categories more often expose the relevant entity, authority, or temporal relation directly, whereas the lower-scoring categories more often require selecting among competing organizational signals. ELE does not isolate relational explicitness as an experimental factor, so we treat this explanation as a hypothesis rather than a causal finding.

Precedent-based exception handling has the lowest mean accuracy, while policy-version reasoning exhibits the widest cross-model spread. Cross-system synthesis remains among the three lowest-scoring dimensions. In these scenarios, multiple records may be individually valid while differing in authority, scope, or time. Retrieving the relevant records is therefore insufficient if the model does not correctly determine which evidence governs the decision. This result motivates separating evidence retrieval from decision-level reconciliation when evaluating enterprise systems.

Model rankings also vary across splits and reasoning dimensions. The Challenge leader differs from the overall leader, and several models show uneven category profiles. These results caution against treating a single aggregate score as a complete description of model behavior on enterprise decisions. Evaluation at the level of the intended organizational relation can reveal weaknesses that are hidden by the benchmark average.

The counterfactual set provides a complementary diagnostic by changing a single fact that should alter the correct organizational decision. Per-scenario accuracy remains high for the strongest model

but varies materially across the evaluated systems. Pair-level required-flip measures would test this behavior more directly and are a natural extension of the present analysis.

ELE measures decision selection rather than end-to-end task execution. A model can select the correct organizational action and still fail to execute it in software, while an agent can complete a workflow correctly at the interface level after making the wrong organizational interpretation. Evaluating these stages separately makes it easier to attribute failures to decision reasoning or execution.

8 LIMITATIONS

Sampling frame. ELE is authored from practitioner-informed patterns rather than sampled from a complete census of enterprise decisions. Scores should therefore be interpreted within the benchmark's scenario distribution, not as the fraction of all enterprise work a model can perform.

Scenario abstraction. Self-contained text scenarios remove environmental complexity such as live tool use, incomplete retrieval, evolving stakeholder conversations, permissions, and interface errors. This increases experimental control but may either understate or overstate deployment difficulty.

Single prompting condition. The reported study uses one direct-prompt condition at temperature 0. It does not estimate the best performance obtainable through model-specific prompting, reasoning controls, retrieval, tools, repeated sampling, or organizational scaffolds.

Scoring with an LLM judge. Exact-match scoring is supplemented by a strict LLM-as-judge threshold of 0.9. Although this avoids semantic partial credit for wrong decisions, any LLM judge can introduce evaluation error. Larger blinded human-calibration studies would strengthen the scoring validity.

No practitioner baseline in the reported run. The current results compare models with one another, not with experienced human practitioners under matched test conditions. A human baseline would help distinguish genuinely difficult organizational ambiguity from model-specific failure.

Counterfactual metric. We report per-scenario accuracy on the 90 counterfactual items and pair-level success on the 45 pairs. We do not yet report required-flip consistency, the fraction of pairs on which the model changes its decision in the direction dictated by the modified governing fact; that metric requires aligning each model's decision label across the two members of a pair and is left to a future release.

Model and benchmark drift. Model versions change quickly, and public benchmark material can eventually enter training corpora. Results should therefore be associated with exact model identifiers and benchmark versions, and future evaluations should maintain model-blind or contamination-resistant test material.

Taxonomy overlap. Real decisions often involve more than one ELE dimension. The six labels are analytical categories, not claims of independent cognitive modules.

9 CONCLUSION

Enterprise's Last Exam evaluates a form of reasoning that is not measured directly by benchmarks centered on generic knowledge, artifact production, or software execution. Across 630 model-blind scenarios and seven proprietary and open-weight models, Core accuracy ranges from 77.4% to 97.4%. Performance is strongly dimension-dependent: the evaluated models are most accurate on entity resolution, approval-chain reconstruction, and temporal consistency, and less accurate on precedent-based exception handling, policy-version reasoning, and cross-system synthesis. The 45 counterfactual pairs evaluate the same systems under single decision-critical factual changes. Taken together, the results identify which forms of organization-specific context remain substantial sources of model error and provide a basis for evaluating the decision layer of enterprise AI separately from execution.

Ethics and Privacy

ELE evaluates AI systems rather than employees or individual companies. Scenarios are written as self-contained enterprise cases and are intended to abstract away customer names, personal data, trade secrets, confidential contract terms, security credentials, and employer-identifying facts. The goal is to preserve the organizational relation that makes a decision difficult without reproducing sensitive incidents verbatim.

Organizational norms vary across companies, jurisdictions, and time. ELE gold decisions therefore should not be interpreted as universal legal, financial, HR, compliance, or employment advice. The benchmark measures whether a model follows the organizational context supplied in the scenario. High benchmark accuracy does not justify unsupervised automation of high-stakes enterprise decisions.

Reproducibility Statement

The benchmark composition, model roster, prompting condition, temperature, scoring rule, and aggregate results are specified in Sections 4–6 and Appendix A; exact provider model identifiers and evaluation configuration are provided in the supplementary files. All three reported evaluation sets were constructed without observing target-model performance. The reported run completed without execution or scoring errors.

Data access. The anonymized supplementary material contains all 630 scenario prompts; the complete evaluation and scoring code; the LLM-judge prompt; model and eval configs; and per-item transcripts for all 4410 model-scenario evaluations. Each transcript includes the model response, extracted answer, scoring method, judge score, and judge reasoning where applicable. To preserve contamination resistance, the complete gold-answer set is withheld during anonymous review. We instead provide a SHA-256 manifest committing to the full private answer set, together with the recorded manifest-generation timestamp. This release strategy is consistent with the broader principle in contamination-aware benchmarking of separating reproducibility artifacts from protected ground truth: TRUCE motivates private benchmark evaluation and auditability, while CapBencher demonstrates that reproducible evaluation can be supported without publicly disclosing underlying ground-truth answers (Rajore et al., 2024; Ishida et al., 2026).

Future benchmark versions should preserve versioned split membership and contamination-resistant evaluation material so that longitudinal comparisons remain interpretable.

AI Use Statement

Generative AI tools were used for literature discovery and for manuscript drafting and editing. All outputs were reviewed and validated by the authors before inclusion. The authors verified the reported benchmark results and citations and take responsibility for the final content of the paper, including all claims and artifacts produced with AI assistance.

References

Boisvert, L., Thakkar, M., Gasse, M., Caccia, M., Le Sellier De Chezelles, T., Cappart, Q., Chapados, N., Lacoste, A., & Drouin, A. (2024). WorkArena++: Towards compositional planning and reasoning-based common knowledge work tasks. *Advances in Neural Information Processing Systems 37*.

Center for AI Safety, Scale AI, & HLE Contributors Consortium. (2026). A benchmark of expert-level academic questions to assess AI capabilities. *Nature, 649*, 1139–1146. https://doi.org/10.1038/s41586-025-09962-4.

Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Del Verme, M., Marty, T., Vazquez, D., Chapados, N., & Lacoste, A. (2024). WorkArena: How capable are web agents at solving common knowledge work tasks? *Proceedings of the 41st International Conference on Machine Learning, PMLR 235*, 11642–11662.

Haimes, J., Wenner, C., Thaman, K., Tashev, V., Neo, C., Kran, E., & Schreiber, J. (2024). Benchmark Inflation: Revealing LLM performance gaps using retro-holdouts. arXiv:2410.09247.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *International Conference on Learning Representations*.

Huang, K.-H., Prabhakar, A., Dhawan, S., Mao, Y., Wang, H., Savarese, S., Xiong, C., Laban, P., & Wu, C.-S. (2025). CRMArena: Understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments. *Proceedings of NAACL 2025*, 3830–3850.

Ishida, T., Lodkaew, T., & Yamane, I. (2026). CapBencher: Give Your LLM Benchmark a Built-in Alarm for Test-Set Overfitting. *Forty-third International Conference on Machine Learning*.

Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., Posada Fishman, S., Aljubeh, M., Thacker, P., et al. (2026). GDPval: Evaluating AI model performance on real-world economically valuable tasks. *International Conference on Learning Representations*.

Rajore, T., Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K., & Swaminathan, M. (2024). TRUCE: Private Benchmarking to Prevent Contamination and Improve Comparative Evaluation of LLMs. arXiv:2403.00393.

Vidgen, B., Mann, A., Fennelly, A., Stanly, J. W., Rothman, L., Burstein, M., Benchek, J., Ostrofsky, D., Ravichandran, A., Sur, D., et al. (2026). APEX-Agents. arXiv:2601.14242.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., et al. (2025). LiveBench: A challenging, contamination-limited LLM benchmark. *International Conference on Learning Representations*.

Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2025). τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *International Conference on Learning Representations*.

Zheng, A., Wu, A., & Liu, A. (2026). ENTLORE: A graph-grounded benchmark for latent organizational reasoning in enterprise question answering. arXiv:2608.10679.

A Full Evaluation Configuration and Results

Table A1: Evaluation configuration.

Component	Configuration
Benchmark size	630 scenarios
Core	384 scenarios
Challenge	156 scenarios
Counterfactual	90 scenarios / 45 pairs
Reasoning dimensions	6
Organizational domains	8
Models	7
Prompt condition	Direct
Temperature	0
Scoring	Exact match, then fixed LLM-as-judge; correctness threshold 0.9
Execution/scoring errors	0
Inference provider	AWS Bedrock (Converse API)
Region	us-west-2
Max output tokens	4096
Parallel workers	4
LLM judge model	claude-opus-4-6 (fixed, separate from evaluated models)
Judge correctness threshold	0.9
Resolved by exact match	73.5% of responses
Non-exact (judge path)	26.5% overall; 22.7–32.4% per model

Table A2: Core-set accuracy with 95% Wilson confidence intervals ($N = 384$).

Model	Core accuracy (%)	95% Wilson CI
claude-opus-5	97.4	[95.3, 98.6]
gpt-5.6-terra	92.9	[90.0, 95.1]
qwen3-235b	86.6	[83.0, 89.8]
llama4-maverick	85.0	[81.0, 88.1]
deepseek-v3.2	85.8	[81.8, 88.8]
gemma-3-27b	82.9	[78.7, 86.3]
nova-micro	77.4	[72.9, 81.2]

Table A3: Category-level accuracy (%).

Model	Entity	Precedent	Cross-system	Policy	Approval	Temporal
gpt-5.6-terra	95	81	90	93	95	94
qwen3-235b	93	78	75	80	94	94
llama4-maverick	92	75	75	79	97	93
claude-opus-5	96	92	89	93	98	95
deepseek-v3.2	93	73	73	78	97	92
gemma-3-27b	89	71	73	68	91	92
nova-micro	87	66	67	63	89	86
Mean	92.1	76.6	77.4	79.1	94.4	92.3

B Interpretation of Counterfactual Results

The counterfactual summary reports per-scenario accuracy for the 90 scenarios and pair-level success for the 45 pairs. Per-scenario accuracy compares model performance on counterfactual material; pair success is the stricter quantity, requiring both members of a pair to be answered correctly.

Table B1 reports pair success by model. gpt-5.6-terra and claude-opus-5 each answer both members correctly on 40 of 45 pairs (89%), while nova-micro is lowest at 28 (62%). The gap between per-scenario accuracy and pair success reflects pairs on which a model answers one member but not the other. A complementary required-flip metric, the fraction of pairs on which the model changes its decision in the direction dictated by the modified governing fact, requires per-pair decision-label alignment and is left to a future release.

Table B1: Counterfactual pair-level results over 45 base/counterfactual pairs.

Model	Both members correct (of 45)	Pair success (%)	One member only
gpt-5.6-terra	40	89	5
qwen3-235b	34	76	10
llama4-maverick	33	73	8
claude-opus-5	40	89	5
deepseek-v3.2	31	69	12
gemma-3-27b	30	67	10
nova-micro	28	62	13