AER-bench: A Comprehensive Benchmark for Automatic Economic Research with Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce **AER-bench**, a comprehensive benchmark for evaluating language models on automatic economic research. Unlike existing benchmarks that focus on isolated question-answering, AER-bench assesses models across the full research pipeline through seven task types: Literature Review, Economic Knowledge, Concept Identification, Paper Abstraction, Economic Analysis, Mathematical Modeling, and Model Generation. Built from 191 top journal articles with automatically extracted tasks and 900 manually annotated historical text segments, our benchmark employs a hybrid evaluation framework combining rule-based scoring and LLM-as-judge assessment across multiple dimensions. We evaluate 16 state-of-the-art language models and reveal three key findings: (1) **pure LLMs show unstable performance on knowledge-intensive tasks**, suggesting the need for agentic systems with external knowledge access; (2) **models within the same family exhibit consistent performance patterns**, indicating systematic capability differences between model series; (3) **understanding of historical economic concepts is unstable across models**: several leading systems handle early-20th-century concepts on par with recent ones, while weaker models drop $20\%+$ on early concepts—a gap that emerges almost entirely on long source paragraphs. Our findings highlight critical gaps in language models' economic research capabilities and provide a rigorous testbed for future development. The benchmark, code, and results are available at GitHub to facilitate reproduction and potential data reconstruction.

1 Introduction

Economic research is a primary input to policy, market design, and aggregate welfare [9, 7, 22, 13, 1, 6, 3], yet producing each contribution still demands a labor-intensive pipeline: literature review, data collection and cleaning, specification of theoretical or econometric models, derivation or estimation, robustness checking, and result interpretation. A substantial share of this pipeline is mechanical and repetitive, and a growing literature has begun to ask how generative AI could absorb it [10].

Frontier large language models are now plausible candidates for that role. Web-scale pretraining has given them broad command of economic terminology, canonical models, and econometric techniques [2, 23], and progress on multi-step reasoning, quantitative problem solving, and formal mathematics [25, 11, 8] brings the symbolic and computational manipulations central to economic analysis within reach. Building on foundational paradigms such as ReAct and Toolformer [26, 20], agentic frameworks like OpenClaw [17], Claude Code [4], and Hermes [15] now package these capabilities into reusable, tool-using workflows that browse, execute code, and iterate on intermediate artifacts, while analogous systems have begun to drive scientific and algorithmic discovery end-to-end [12, 16]. A first attempt has already appeared in economics itself: APE (Automatic Policy Evaluation) [21], built on Claude Code, automates empirical policy research, yet on its own self-

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

reported scale still trails top journals by a wide margin—at once demonstrating the promise of automated economic research and underscoring the need for an independent, rigorous benchmark.

Table 1: Comparison of AER-bench with existing economic scenario benchmarks. AER-bench is the first benchmark designed to evaluate the pipeline of automatic economic research.

Benchmark	Data Source	Focus	Modality	# Items	# Task Types
EconGym [14]	Real-world data	Econ Behavior	Simulation	—	25+
GDPVal [18]	Industry reports	Econ Productivity	Text	1,320	9
EconLogicQA [19]	News articles	Econ Reasoning	Text	650	1
FinMR [5]	Web + Experts	Fin Reasoning	Text + Image	3,700	15
AER-bench (Ours)	Journals + Experts	Econ Research	Text	1,629	7

General-purpose academic benchmarks such as MMLU-Pro [24] include an economics subject, but each subject contributes only a few hundred multiple-choice items and cannot probe the research-grade capabilities we are interested in. Several economics-focused benchmarks[1] have therefore been proposed (Table 1). EconGym [14] is a scalable multi-agent simulation testbed in which LLMs play the role of households, firms, or governments across more than twenty real-data-driven economic tasks, evaluating decision behavior rather than research output. GDPVal [18] curates 1,320 tasks across nine GDP-relevant industries from industry reports to measure whether AI can substitute for economically valuable human labor, again outside the academic pipeline. EconLogicQA [19] distills 650 multi-event sequencing questions from economic news articles, targeting the narrow capability of causal and temporal reasoning. FinMR [5] provides over 3,700 multimodal questions on financial-analyst-level chart and formula interpretation, likewise sourced from non-academic web material rather than economics research. Each benchmark is valuable within its own scope—behavioral simulation, productive substitution, event reasoning, and financial analysis, respectively—yet none draws its data from top-journal research articles, and consequently none evaluates the core capabilities that economic research relies on, such as domain knowledge, mathematical modeling, and economic analysis of empirical or theoretical results. This gap remains unaddressed.

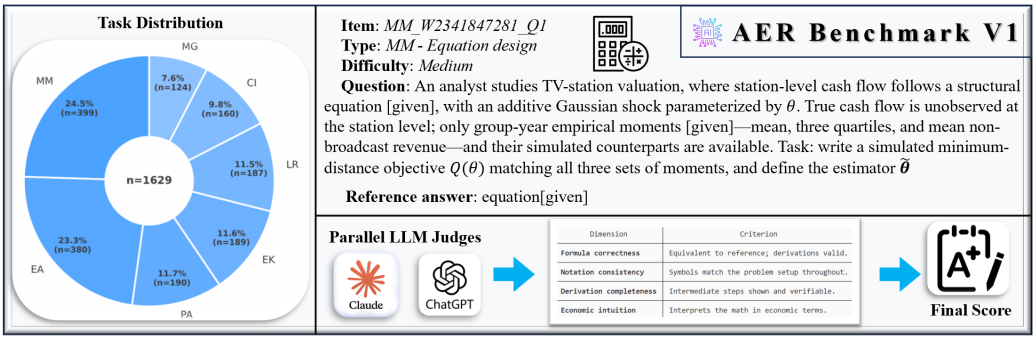

Figure 1: Overview of the AER benchmark. **Left**: a pie chart showing the distribution and proportion of tasks across the seven categories; **middle**: a representative task example; **bottom**: the evaluation pipeline, in which `gpt-5.2` and `claude-4.6-sonnet` independently score each response against a shared rubric, and their judgments are aggregated into a final score.

To address this gap, we introduce **AER-bench**[2], the first comprehensive benchmark for systematically and objectively evaluating language models on automatic economic research. AER-bench combines two complementary data sources: (i) 191 articles drawn from top-5 economics journals—the *American Economic Review* and *Econometrica*—representing the modern frontier of the field; and (ii) roughly 900 text segments from pre-1950 economics literature, which support the analysis of how language models handle historical economic discourse. Tasks are constructed through two complementary pipelines (§2.2): an LLM-driven multi-round extraction pipeline that turns paper

[1]We discuss only a few representative economics-related benchmarks here; additional benchmarks covering economic and financial scenarios exist but are omitted for brevity.

[2]An overview of our benchmark is shown in Figure 1, where the representative task example illustrates that our tasks differ fundamentally from those in previous benchmarks.

PDFs into self-contained items, and a manual annotation effort by economics doctoral students that distils the historical segments into 160 concept-identification tasks across eight conceptual dimensions. The result is a benchmark of 1,629 evaluation items organized into seven task types—EK, MM, PA, LR, MG, EA, and CI—spanning capabilities from foundational knowledge recall and symbolic manipulation to higher-order modeling and analysis (§2.1). To score these heterogeneous tasks reliably, AER-bench adopts a hybrid evaluation framework (§2.3): rule-based scoring for EK and CI, and multi-dimensional rubric-based LLM-as-judge scoring averaged over two independent judges (`claude-4.6-sonnet` and `gpt-5.2`) for the remaining five tasks. We use AER-bench to evaluate 16 frontier LLMs (§3); the full ranking and per-task breakdowns are reported in §4.

In summary, AER-bench is, to our knowledge, the first comprehensive benchmark specifically designed to evaluate language models on the full pipeline of automatic economic research, combining a workflow-aligned task taxonomy, top-journal and pre-1950 sources, and a hybrid rule-based and dual-judge evaluation framework. Beyond the benchmark itself, the underlying construction methodology—an LLM-driven multi-round extraction pipeline coupled with expert annotation—is in itself a reusable asset for continuously extending the benchmark, constructing training data, and adapting the approach to other research disciplines. Across 16 frontier LLMs we observe pervasive shortcomings, most notably substantial instability on knowledge-intensive tasks, which we interpret as evidence that agentic, retrieval-augmented architectures are a necessary next step toward reliable automated economic research. Along the way, AER-bench surfaces several further phenomena worth attention—including consistent within-family capability patterns and a pronounced historical concept drift (a "modernity bias") on pre-1950 economic concepts—which we examine in detail in §4.

2 Benchmark Design

To systematically evaluate the capability of Language Models in automating economic research, we introduce AER-bench. Unlike generic benchmarks (e.g. MMLU) that provide only shallow coverage of economic scenarios, AER-bench is grounded in the authentic research pipeline, assessing models across both deterministic knowledge retrieval and open-ended generative analysis.

2.1 Task Taxonomy

AER-bench deconstructs the economic research workflow into seven distinct task categories. Each task targets a specific phase of the research lifecycle and is evaluated along fine-grained dimensions:

- **Literature Review (LR):** Assesses the model's ability to reconstruct the argument structure of an introduction and appropriately cite prior literature. It is evaluated via a dual mechanism: citation recall (capturing the core reference network) and a paper-specific writing rubric.
- **Economic Knowledge (EK):** A rigorous QA task comprising single-choice, true/false, fill-in-the-blank, and ordering questions. It tests the model's deep comprehension of paper-specific methodologies, findings, and institutional contexts.
- **Concept Identification (CI):** A robust binary classification task requiring models to map text chunks to eight structural economic properties: relevance, mathematical formalization, uniqueness, stability, existence, equilibrium scene, time characteristic, and time horizon.
- **Paper Abstraction (PA):** The model must synthesize a concise abstract from a condensed paper body. Performance is evaluated across four dimensions: coverage of atomic semantic propositions, factual accuracy, conciseness, and academic style.
- **Economic Analysis (EA):** Given uninterpreted empirical or theoretical results, the model generates a qualitative analysis. It is scored across six dimensions: accuracy, economic interpretation, mechanism, theory connection, policy implication, and writing quality.
- **Mathematical Modeling (MM):** Given a textual economic scenario and notation constraints, the model must perform formal derivations. Outputs are judged on four dimensions: correctness, notation consistency, derivation completeness, and economic intuition.
- **Model Generation (MG):** Evaluates formal construction of an economic model from a de-symbolized narrative. Scoring uses a granular element rubric checking for presence and correctness of sectors, objective functions, constraints, and equilibrium concepts.

2.2 Data Collection & Generation Pipeline

Data Sources To ensure the benchmark is both comprehensive, AER-bench combines an automated extraction pipeline for modern research with meticulous manual annotations for historical literature. The dataset is built upon two distinct corpora:

- *Modern Frontiers (191 articles):* Sourced from public, Open Access[3] versions of papers published in 2025 in top-tier journals (*American Economic Review*, *Econometrica*), guaranteeing that the benchmark challenges models with the most recent methodological innovations while strictly adhering to copyright and open-science protocols.
- *Historical Fragments (900 segments):* Curated from multilingual text segments of canonical historical literature (1900–1950) to evaluate models on historical concept drift and early economic paradigms, which presents a challenging task due to archaic terminology and evolving theoretical frameworks.

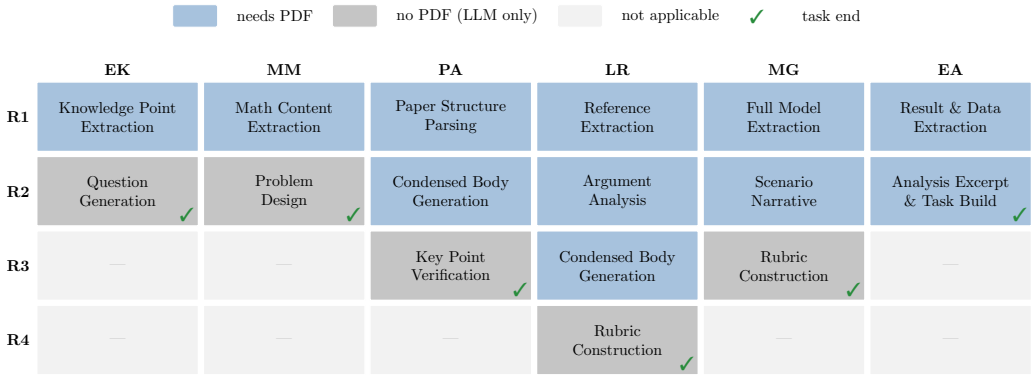

Figure 2: Auto-acquisition pipelines for the six LLM-judged tasks. Each column is a task type and each row is a generation round; cell labels name the per-round operation. Blue cells consume the source PDF, gray cells operate from prior-round outputs alone (LLM-only), and the green check marks each task's terminal round. Pipeline depth ranges from 2 to 4 rounds.

Task Generation Pipeline For the modern corpus, we employ a multi-round, LLM-assisted generation pipeline that begins with structural PDF parsing to extract raw text and equations, followed by multiple rounds of information extraction, desymbolization, and task construction with automated quality checks to ensure self-contained tasks free of structural errors[4]. Figure 2 summarizes the round-by-round operations across all six modern-corpus pipelines, illustrating which rounds consume the source PDF and where each pipeline terminates; pipeline depth ranges from two rounds (EK, MM, EA) to four rounds (LR), reflecting the differing structural complexity of each task. In contrast, the historical fragments used for the Concept Identification (CI) task were strictly manually annotated by expert annotators[5] with advanced degrees in economics, who read each fragment and performed binary classification across eight predefined conceptual fields, providing gold-standard labels that form a rigorous testbed for evaluating AI comprehension of early, non-formalized economic concepts.

2.3 Evaluation Framework

AER-bench utilizes a hybrid evaluation framework that balances absolute objectivity with nuanced semantic assessment, adapting its scoring strategy to the nature of each task type.

[3]For articles not available under Open Access licenses from journals, we substitute with legally accessible working paper versions from NBER, arXiv, or authors' personal homepages to ensure data legality.

[4]See Appendix G for detailed prompt templates, output formats, and scoring methods for each task.

[5]All annotators are economics doctoral students who have passed their qualifying examinations and language requirements. Moreover, we only select and retain data with consistent annotations to ensure the validity of the task answers.

Rule-Based Scoring For tasks with deterministic answers (EK and CI), we employ strict rule-based grading. CI uses categorical exact matching, while EK incorporates unidirectional word-boundary regular expressions for fill-in-the-blank items and normalized Kendall's τ distance for ordering tasks.

LLM-as-Judge For open-ended generative tasks[6] (MM, PA, LR, MG, EA), we adopt an LLM-as-Judge paradigm. To prevent holistic scoring biases, each judge model evaluates outputs across multiple isolated dimensions with detailed rubrics, producing a weighted dimension-level score per task that captures distinct aspects of quality. To further enhance robustness, we employ multiple judge models in parallel, mitigating individual model biases. Section 4.2 systematically compares the scoring differences between two parallel judges to validate this approach.

Score Aggregation For each item, dimension scores are first aggregated into a task score via configurable intra-task weights, then scores from multiple judges are arithmetically averaged to produce a consolidated item score. Overall benchmark scores are computed by weighted summation across task types, with specific weight values detailed in Section 3.

3 Experimental Setup

Evaluated Models & Inference Details. We evaluate 16 state-of-the-art Large Language Models to establish a comprehensive baseline for automated economic research. The cohort includes the OpenAI GPT family (`gpt-5.4`, `gpt-5.2`, `gpt-5.1`, `gpt-5`, `gpt-5-mini`, `gpt-4o`), the Anthropic Claude series (`claude-4.6-opus`, `claude-4.6-sonnet`, `claude-3.7-sonnet`), and the Google Gemini lineage (`gemini-3.1-pro`, `gemini-3-flash`, `gemini-2.5-flash`). Furthermore, we incorporate leading open-weight and regional models: `kimi-k2.5`, `glm-5.1`, `doubao-seed-1.6`, and `qwen3-vl-32b-instruct`. Note that all inferences are executed under a zero-shot setting.

Table 2: Dimension-level scoring weights for each task type. Dimensions are weighted according to their importance for evaluating model performance, with weights normalized to sum to 1 within each task. Detailed rubric definitions for each dimension are provided in Appendix E.

Task	**EK**	**MM**	**PA**	**LR**	**MG**	**EA**	**CI**
# Dims	4	4	4	2	Multiple	6	8
Weights	Equal	4:2:3:1	4:4:3:1	3:2	Equal	4:4:4:3:3:2	Equal

Judge Configuration & Score Aggregation. For the hybrid evaluation framework introduced in Section 2.3, we instantiate a programmatic evaluator for deterministic tasks (EK and CI). For the five open-ended generative tasks, we deploy `claude-4.6-sonnet` and `gpt-5.2` as independent LLM-as-Judges to mitigate individual evaluator bias.

To formalize the multi-judge merging strategy outlined previously, let J be the set of judges evaluating model m on task t. The task-level score is computed as the arithmetic mean of the judges' outputs:

$$S_{m,t} = \frac{1}{|J|} \sum_{j \in J} s^j_{m,t} \tag{1}$$

The overall benchmark score S_m is then calculated as the weighted sum of the task-level scores. In AER-bench, all seven task categories are assigned equal importance ($w_t = 1/7$):

$$S_m = \sum_{t \in T} w_t \cdot S_{m,t}, \quad \sum_{t \in T} w_t = 1 \tag{2}$$

Within each multi-dimensional task, sub-scores are aggregated according to predefined intra-task dimension weights, which reflect the relative importance of specific economic reasoning facets. These exact dimensional configurations are detailed in Table 2.

[6]Note that the Literature Review (LR) task employs both an LLM-as-Judge rubric and a rule-based citation recall score, which are combined via weighted aggregation.

4 Results

4.1 Model Performance

Table 3 presents the overall benchmark ranking across 16 evaluated language models. `gpt-5.4` achieves the highest overall score at 75.0%, excelling in Mathematical Modeling (92.8%), Model Generation (85.8%), and Economic Analysis (81.6%); however, a paired bootstrap test shows this lead is statistically indistinguishable from `claude-4.6-opus` (74.0%, $p = 0.092$), placing the two models in a statistical tie at the top. OpenAI models nonetheless exhibit notable version instability: `gpt-4o` scores only 54.0% with particularly weak Economic Knowledge (25.6%). In contrast, Anthropic's Claude models demonstrate remarkably stable performance, with `claude-4.6-opus` (74.0%), `claude-3.7-sonnet` (73.7%), and `claude-4.6-sonnet` (73.4%) clustered within 0.6 percentage points, and achieve the highest Concept Identification scores (72.7–72.9%), substantially outperforming `gpt-5.4` (58.9%). Google's `gemini-3.1-pro` (71.6%) exhibits task-specific dominance, achieving state-of-the-art performance on Economic Knowledge (76.3%) and Paper Abstraction (74.3%) despite trailing in overall ranking, possibly reflecting targeted training on academic literature. `kimi-k2.5` (69.1%) demonstrates that open-weight models are approaching closed-source performance on domain-specific benchmarks. Critically, within-family performance generally correlates with model version (`gpt-5.4` > `gpt-5.2` > `gpt-5` > `gpt-5.1`), validating that AER-bench effectively captures incremental capability improvements. Per-task 95% bootstrap confidence intervals for all 16 models are reported in Appendix A (Table 4).

Table 3: Overall benchmark scores (in %) with item-level 95% bootstrap confidence intervals on the Overall column. Per-task columns show point estimates; bold values mark the best score in each column. Models are grouped by provider and sorted by Overall within each group.

Provider	Model	Overall	EK	MM	PA	LR	MG	EA	CI
OpenAI	gpt-5.4	**75.0** [74.0, 76.0]	73.9	**92.8**	76.2	**56.0**	**85.8**	**81.6**	58.9
	gpt-5.2	73.0 [72.0, 74.0]	70.8	91.5	76.4	50.3	83.6	78.2	60.5
	gpt-5	71.0 [69.9, 72.0]	74.2	91.8	73.5	47.7	84.2	78.5	46.9
	gpt-5.1	70.0 [69.0, 71.0]	67.5	89.7	75.6	48.1	82.8	75.0	51.4
	gpt-5-mini	69.0 [68.1, 70.0]	72.2	90.8	71.1	48.1	79.2	76.8	45.1
	gpt-4o	54.0 [53.1, 55.0]	25.6	78.4	72.1	39.5	45.7	64.6	52.3
Anthropic	claude-4.6-opus	74.0 [73.1, 74.9]	71.5	89.4	76.1	52.5	78.1	77.7	72.7
	claude-3.7-sonnet	73.7 [72.8, 74.6]	67.0	88.6	76.8	50.7	81.7	78.2	**73.0**
	claude-4.6-sonnet	73.4 [72.5, 74.3]	66.7	88.8	76.3	50.6	81.8	76.9	72.9
Google	gemini-3.1-pro	71.6 [70.6, 72.5]	**76.3**	89.3	74.3	47.0	70.0	72.0	72.0
	gemini-3-flash	65.8 [64.9, 66.8]	74.8	86.5	**77.0**	46.2	62.4	72.3	41.6
	gemini-2.5-flash	64.9 [63.9, 65.9]	69.4	88.0	76.1	41.4	68.0	67.1	44.1
Others	kimi-k2.5	69.1 [68.0, 70.1]	68.9	87.4	73.5	47.3	71.9	76.3	58.3
	glm-5.1	65.5 [64.4, 66.6]	57.9	85.4	75.8	49.0	69.5	74.8	46.3
	doubao-seed-1.6	64.8 [63.8, 65.8]	66.0	84.9	74.7	42.2	61.5	72.7	51.4
	qwen3-vl-32b-instruct	61.3 [60.3, 62.3]	60.9	82.0	73.5	39.1	59.4	66.2	48.0

To preempt concerns that the ranking above might be an artifact of sampling noise, we additionally conduct pairwise paired-bootstrap significance tests on the Overall score for all $\binom{16}{2}$ model pairs (Appendix B, Table 6): every cross-tier comparison rejects equality at $p < 0.001$, while a small number of within-cluster pairs—most notably `gpt-5.4` vs. `claude-4.6-opus`—remain statistically tied, providing a conservative significance frame for every claim we make about model rankings.

Figure 3 illustrates the variation in task difficulty and discriminative power across the benchmark. Literature Review presents the lowest median score ($\sim$40%) with tight clustering, indicating that all models struggle uniformly with citation network reconstruction and argument structure recovery. Conversely, Mathematical Modeling exhibits high scores (median near 90%) with limited spread, suggesting that symbolic manipulation capabilities have largely converged across modern model families. Tasks like Economic Knowledge and Model Generation demonstrate the widest score ranges, serving as the primary differentiators between average and frontier AI systems in economic research contexts. See Appendix F for detailed performance analysis by task dimension.

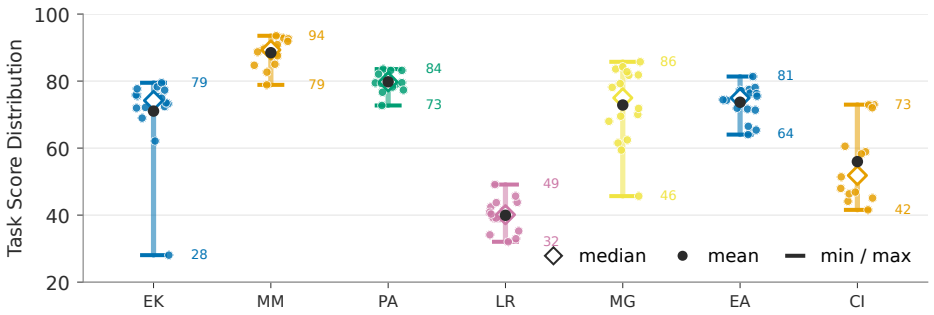

Figure 3: Task score distribution across all 16 evaluated models. Each vertical line represents the min-max range of model scores for a given task, with individual model scores shown as colored dots. White diamonds indicate median scores, and gray circles indicate mean scores. The figure reveals substantial performance variation across tasks.

4.2 LLM-as-a-Judge Quality

To validate the multi-judge evaluation framework detailed in Section 3, we analyze the scoring consistency between two independent judge models: `claude-4.6-sonnet` and `gpt-5.2`. Figure 4 shows high overall agreement, with a mean absolute difference of 2.6 percentage points. This tight clustering confirms that our dimension-level, criteria-bound rubrics effectively constrain the subjectivity typical of LLM evaluators, producing reliable assessments across different judge architectures.

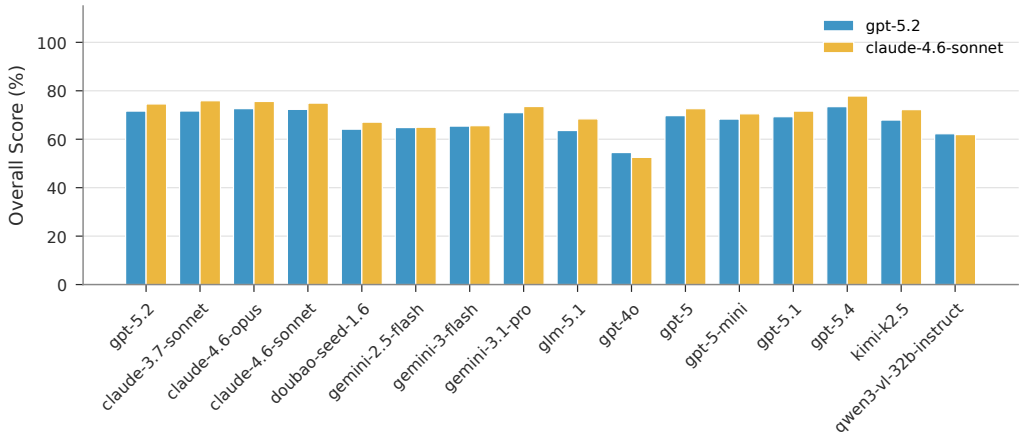

Figure 4: Judge agreement on overall benchmark scores across 16 models. Scores from two independent LLM judges (`claude-4.6-sonnet` and `gpt-5.2`) show strong consistency, with a mean absolute difference of 2.6 percentage points, validating the robustness of our rubric-based evaluation framework.

While overall agreement is strong, Figure 5 reveals systematic task-level scoring preferences between the judges. `claude-4.6-sonnet` assigns higher scores on Mathematical Modeling (+3.1 pp) and Model Generation (+13.0 pp), whereas `gpt-5.2` scores Paper Abstraction higher (-3.4 pp). To test whether these task-level offsets translate into a structural *self-preference bias*—each judge inflating scores for its own family—we computed the per-item Claude–minus–GPT score gap $\Delta_{m,i}$ for every examinee m, applied a paired Wilcoxon signed-rank test per model, and then compared the per-model mean Δ across families with a two-sided Mann–Whitney U test. Anthropic candidates do receive a larger "Claude-judge bonus" than OpenAI candidates (mean $\Delta = +2.90$ pp vs. $+1.41$ pp; "Other" families $+0.82$ pp), but the family-level difference falls short of conventional significance ($p = 0.095$, $n = 3$ vs. $n = 6$). The directional effect is real but small relative to the cross-tier gaps documented in Table 3, and arithmetic averaging across the two independent judges substantially attenuates it; per-model results and the full statistical analysis are reported in Appendix C.

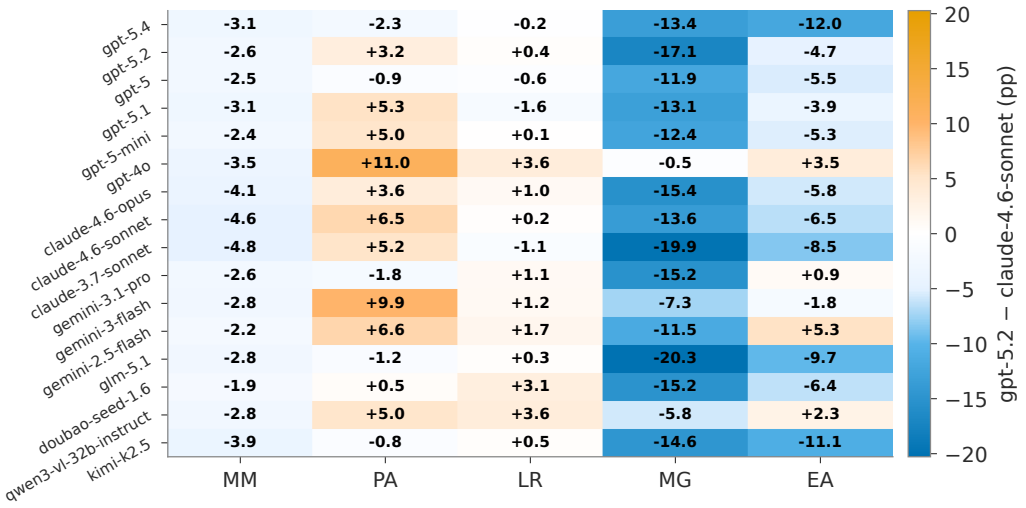

Figure 5: Task-level scoring differences of two judges. Heatmap shows the score difference (`claude-4.6-sonnet` minus `gpt-5.2`) across five generative tasks, with blue indicating higher scores from Claude and orange indicating higher scores from `gpt-5.2`. Models are grouped by provider and sorted by overall performance within each group.

4.3 Key Findings

Based on the quantitative results, we highlight three core findings.

Finding 1: Performance Volatility in Knowledge-Intensive Tasks Demands Agentic Solutions. Pure language models exhibit severe instability when tasked with retrieving specific literature facts or constructing citation networks. This is evident in the extremely wide spread of Economic Knowledge (EK) scores (ranging from 28.1% to 79.5%) and the universally low performance in Literature Review (LR) shown in Figure 3. These results suggest that relying solely on an LLM's parametric memory is insufficient for rigorous academic writing. Achieving expert-level reliability in these phases will likely require integrating pure LLMs with agentic systems.

Finding 2: Distinct Family-Level Capability Fingerprints. Beyond version-level scaling, AER-bench reveals task-specific signatures that distinguish model families. All five GPT-5 variants occupy the top five positions on MM (90.9–93.5%), leading the best non-GPT-5 model by 3–4 points; the 3 Claude models cluster tightly at 72.7–73.0% on CI—over 12 points above `gpt-5.2`—while staying within 0.5% overall, indicating a stable shared profile rather than checkpoint-level idiosyncrasy; and Gemini models trail GPT-5 and Claude by 5–15 points on LR, MG, and EA despite being competitive on EK and PA. These cross-family patterns are systematic rather than random, showing that AER-bench can diagnose the relative strengths of entire series, not merely rank isolated models.

Finding 3: A Model-Stratified, Length-Conditional "Modernity Bias." At face value, our longitudinal Concept Identification (CI) task suggests a pervasive "modernity bias": aggregate scores are $\sim$5 pp higher on 1930–1950 than on 1900–1929 chunks (Figure 6), led by the `equilibrium_scene` dimension. Two robustness checks (Appendix D, Tables 9–8) sharpen this picture. After partialling out source-chunk length (OLS, model FE, item-clustered SE), the $\text{Era}_3 - \text{Era}_2$ coefficient shrinks to $+2.75$ pp ($p = 0.44$); the gap lives entirely in the longest two length quartiles (Q3 $+9.2$, Q4 $+12.7$ pp), while the shortest two *reverse* (Era_2 leads by 1.8–2.1 pp). A per-decade $\times$ per-model breakdown then reveals strong heterogeneity: the three Claude models, `gemini-3.1-pro`, `gpt-5.2`, and `kimi-k2.5` score *higher* on 1900–1909 than on 1940–1950 (e.g., `claude-4.6-sonnet` 86.2 vs. 64.2%), whereas `gpt-5.4`, `gpt-5`, `gpt-5-mini`, both Gemini Flash variants, and `glm-5.1` show the canonical 20+ pp drop on the earliest decade. The honest characterisation is therefore a *model-stratified, length-conditional* bias: real for specific models on longer paragraphs, but closed—and in some cases inverted—by frontier Anthropic, Gemini-3.1, and GPT-5.2 systems. The pattern points to

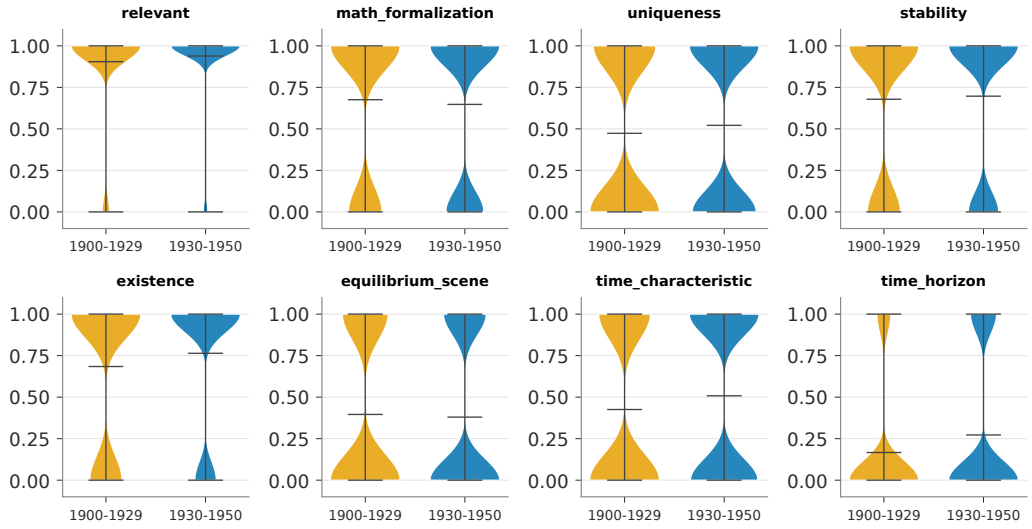

Figure 6: Concept identification performance across historical periods. Violin plots show the distribution of 16 model scores on each of the 8 concept dimensions, comparing publications from 1900–1929 (orange) versus 1930–1950 (blue). Models exhibit systematic performance gaps on earlier texts, particularly for `equilibrium_scene` and `existence` dimensions.

representational gaps in pre-training corpus composition rather than uniform conceptual drift, and reframes the CI task as a diagnostic for *which* models still carry this blind spot.

5 Limitations

While AER-bench establishes a rigorous evaluation framework, it presents several inherent limitations. First, the automated task generation pipeline is computationally expensive. Processing full-length papers through multi-turn interactions requires top-tier models, currently costing approximately $6 per paper. Consequently, we deliberately prioritized high-fidelity task extraction over massive scale. Second, the reliance on PDF parsing occasionally introduces formatting artifacts that can disrupt the extraction of complex mathematical notations. Third, although our dimension-level rubrics mitigate subjective biases, the LLM-as-a-Judge paradigm may still harbor residual architectural preferences. Finally, the modern corpus consists predominantly of English-language journals, potentially masking disparities in models' multilingual economic reasoning capabilities. Additionally, the benchmark does not include a human expert baseline, a decision motivated by several practical and conceptual challenges discussed in the Appendix H.

6 Conclusion

This paper introduces AER-bench, the first comprehensive benchmark designed to evaluate LLMs across the full lifecycle of economic research. Moving beyond simple question-answering paradigms, AER-bench assesses 16 state-of-the-art models on a rigorous, multi-dimensional taxonomy spanning seven distinct tasks—from deterministic knowledge retrieval and conceptual identification to the generation of formal mathematical models and qualitative empirical analyses.

Our evaluations reveal that while frontier models excel in symbolic manipulation and text summarization, they exhibit high volatility in knowledge-intensive tasks like citation network reconstruction. We also uncover a systematic "modernity bias" wherein models struggle to interpret early historical economic concepts. These findings highlight the necessity of transitioning from pure, parametrically-reliant LLMs to agentic, retrieval-augmented systems. Ultimately, AER-bench provides a robust and systematic evaluation yardstick that will guide the development of the next generation of AI systems capable of autonomously conducting, interpreting, and advancing rigorous economic research.

References

[1] Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401, 2001.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

[4] Anthropic. Claude code. `https://github.com/anthropics/claude-code`, 2025. Accessed: 2026-05.

[5] Shuangyan Deng, Haizhou Peng, Jiachen Xu, Rui Mao, Ciprian Doru Giurcaneanu, and Jiamou Liu. *FinMR: A Knowledge-Intensive Multimodal Benchmark for Advanced Financial Reasoning*, page 168–176. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400722202. URL `https://doi.org/10.1145/3768292.3770365`.

[6] Esther Duflo and Abhijit Banerjee. *Poor economics*, volume 619. PublicAffairs New York, 2011.

[7] Milton Friedman et al. The methodology of positive economics. *Essays in positive economics*, 3(3):145–178, 1953.

[8] Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, pages 1–3, 2025.

[9] John Maynard Keynes. The general theory of employment. *The quarterly journal of economics*, 51(2):209–223, 1937.

[10] Anton Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

[11] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[12] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[13] Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

[14] Qirui Mi, Qipeng Yang, Zijun Fan, Wentian Fan, Heyang Ma, Chengdong Ma, Siyu Xia, Bo An, Jun Wang, and Haifeng Zhang. Econgym: A scalable ai testbed with diverse economic tasks. *arXiv preprint arXiv:2506.12110*, 2025.

[15] Nous Research. Hermes agent. `https://github.com/nousresearch/hermes-agent`, 2025. Accessed: 2026-05.

[16] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[17] OpenClaw Contributors. OpenClaw: An open-source agentic framework. `https://github.com/openclaw/openclaw`, 2025. Accessed: 2026-05.

[18] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

[19] Yinzhu Quan and Zefang Liu. EconLogicQA: A question-answering benchmark for evaluating large language models in economic sequential reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2273–2282, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.125. URL https://aclanthology.org/2024.findings-emnlp.125/.

[20] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36: 68539–68551, 2023.

[21] Social Catalyst Lab. APE: Automatic Policy Evaluation. https://ape.socialcatalystlab.org/, 2025. Accessed: 2026-05.

[22] Robert M Solow. Technical change and the aggregate production function. *The review of Economics and Statistics*, 39(3):312–320, 1957.

[23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[26] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

A Full Per-Task Ranking with Bootstrap Confidence Intervals

To complement the main-text ranking (Table 3), we report in Table 4 the full per-task results for all 16 evaluated models, with item-level 95% bootstrap confidence intervals on both the Overall score and every individual task. Confidence intervals are obtained from $B = 10{,}000$ bootstrap resamples stratified within each task; the same resample indices are shared across models so that pairwise comparisons preserve per-item correlation. For non-CI tasks, the per-item score is the arithmetic mean of the `claude-4.6-sonnet` and `gpt-5.2` judges before resampling; the CI task uses the rule-based judge.

No single model dominates all dimensions. While `gpt-5.4` attains the column-wise maximum on five of the eight columns (Overall, MM, LR, MG, EA), it is decisively beaten on the remaining three: `gemini-3.1-pro` leads on EK (76.3 vs. 73.9), `gemini-3-flash` leads on PA (77.0 vs. 76.2), and the Anthropic family leads on CI by a wide margin (`claude-3.7-sonnet` 73.0; `claude-4.6-sonnet` 72.9; `claude-4.6-opus` 72.7) versus `gpt-5.4` at 58.9. The CI column in particular splits cleanly along provider lines: every Anthropic model exceeds 72%, whereas the OpenAI family except `gpt-4o` (52.3) sits between 45.1 and 60.5. Combined with the comparable ranking of `gemini-3.1-pro` (72.0) on the same task, this pattern suggests that performance on early-20th-century economic-history concepts depends more on pre-training corpus composition than on raw scale.

Confidence-interval width tracks item count and judge type. CI widths are far from uniform across tasks. The two largest tasks, EA ($n = 380$) and MM ($n = 399$), produce the tightest intervals: half-widths are typically ≤ 0.6 pp and ≤ 0.8 pp respectively (e.g., `gpt-5.4` on EA has $[81.0, 82.2]$). The CI task is the opposite extreme, with half-widths near 5–6 pp on every model (e.g., `claude-4.6-opus` $[67.3, 77.8]$), reflecting both its smaller sample ($n = 160$) and the higher per-item variance of multi-field rule-based scoring. LR ($n = 187$) and MG ($n = 124$) show intermediate widths around 1–3 pp. These differences are quantitative reminders that gaps of 1–2 pp on CI should not be over-interpreted, whereas similar-sized gaps on EA or MM are typically well outside the bootstrap noise floor.

Anthropic models occupy a tight cluster. The three Claude models lie within 0.6 pp on Overall (74.0, 73.7, 73.4) and their per-task CIs overlap extensively on every column. Their largest within-family spread is on EK (66.7–71.5), where `claude-4.6-opus` clearly leads, and on MG (78.1–81.8), where the two Sonnet variants slightly exceed Opus—suggesting the Sonnet line carries a marginal advantage on procedural model construction. By contrast, the OpenAI line shows a near-monotonic version-vs-score relationship on Overall (`gpt-5.4` > `gpt-5.2` > `gpt-5` > `gpt-5.1` > `gpt-5-mini` > `gpt-4o`), with `gpt-4o` a clear outlier on EK at 25.6, more than 25 pp below every other model and consistent with its older training cutoff.

Where the rankings are statistically separable. Reading the bootstrap intervals horizontally provides an informal visual significance test on the Overall column: the top two rows (`gpt-5.4` $[74.0, 76.0]$ and `claude-4.6-opus` $[73.1, 74.9]$) overlap by roughly 1 pp, consistent with the paired-bootstrap p-value of 0.092 reported in the main text. The next three rows (`claude-3.7-sonnet`, `claude-4.6-sonnet`, `gpt-5.2`) form a tightly overlapping cluster between 72.5 and 74.6. After `gemini-3.1-pro` the intervals start to separate cleanly, with `gpt-4o` ($[53.1, 55.0]$) sitting more than 6 pp below the next-lowest model. Per-task intervals tell the same story at finer resolution: many cross-provider gaps on EK and CI are highly significant (no overlap at all), while LR and PA gaps are typically within the noise floor.

Table 4: Per-task and overall scores (in %) with item-level 95% bootstrap confidence intervals ($B = 10{,}000$ resamples, stratified within task). Each cell stacks the point estimate above its $[\mathrm{ci}_{\mathrm{lo}}, \mathrm{ci}_{\mathrm{hi}}]$ interval. Bold marks the column-wise maximum. Models are grouped by provider and sorted by Overall within each group.

Provider	Model	Overall	EK	MM	PA	LR	MG	EA	CI
OpenAI	gpt-5.4	**75.0** [74.0, 76.0]	73.9 [72.4, 75.4]	**92.8** [92.0, 93.5]	76.2 [73.9, 78.5]	**56.0** [54.7, 57.3]	**85.8** [83.2, 88.1]	**81.6** [81.0, 82.2]	58.9 [53.0, 64.8]
	gpt-5.2	73.0 [72.0, 74.0]	70.8 [69.3, 72.2]	91.5 [90.8, 92.3]	76.4 [74.7, 78.0]	50.3 [49.2, 51.3]	83.6 [81.2, 85.8]	78.2 [77.7, 78.6]	60.5 [54.5, 66.6]
	gpt-5	71.0 [69.9, 72.0]	74.2 [72.6, 75.8]	91.8 [91.0, 92.6]	73.5 [71.2, 75.6]	47.7 [46.9, 48.5]	84.2 [81.6, 86.7]	78.5 [78.0, 78.9]	46.9 [40.9, 52.9]
	gpt-5.1	70.0 [69.0, 71.0]	67.5 [65.9, 69.0]	89.7 [88.8, 90.6]	75.6 [73.6, 77.5]	48.1 [47.2, 49.0]	82.8 [80.6, 84.8]	75.0 [74.4, 75.6]	51.4 [45.5, 57.3]
	gpt-5-mini	69.0 [68.1, 70.0]	72.2 [70.7, 73.7]	90.8 [89.8, 91.7]	71.1 [69.0, 73.1]	48.1 [47.0, 49.2]	79.2 [76.8, 81.5]	76.8 [76.2, 77.3]	45.1 [39.3, 50.9]
	gpt-4o	54.0 [53.1, 55.0]	25.6 [23.2, 28.2]	78.4 [77.0, 79.8]	72.1 [70.9, 73.1]	39.5 [38.7, 40.4]	45.7 [43.7, 47.7]	64.6 [63.9, 65.4]	52.3 [46.3, 58.0]
Anthropic	claude-4.6-opus	74.0 [73.1, 74.9]	71.5 [70.0, 73.0]	89.4 [88.4, 90.5]	76.1 [74.3, 78.0]	52.5 [51.4, 53.6]	78.1 [75.4, 80.7]	77.7 [77.1, 78.2]	72.7 [67.3, 77.8]
	claude-3.7-sonnet	73.7 [72.8, 74.6]	67.0 [65.4, 68.5]	88.6 [87.6, 89.7]	76.8 [75.1, 78.4]	50.7 [49.5, 51.8]	81.7 [79.5, 83.8]	78.2 [77.6, 78.7]	**73.0** [67.5, 78.2]
	claude-4.6-sonnet	73.4 [72.5, 74.3]	66.7 [65.1, 68.3]	88.8 [87.7, 89.8]	76.3 [74.5, 78.0]	50.6 [49.4, 51.6]	81.8 [79.4, 84.1]	76.9 [76.3, 77.5]	72.9 [67.6, 78.0]
Google	gemini-3.1-pro	71.6 [70.6, 72.5]	**76.3** [75.0, 77.6]	89.3 [88.3, 90.3]	74.3 [71.5, 77.0]	47.0 [46.2, 47.9]	70.0 [67.3, 72.6]	72.0 [71.5, 72.6]	72.0 [67.1, 77.0]
	gemini-3-flash	65.8 [64.9, 66.8]	74.8 [73.4, 76.2]	86.5 [85.5, 87.6]	**77.0** [75.7, 78.1]	46.2 [45.2, 47.2]	62.4 [60.4, 64.4]	72.3 [71.6, 72.9]	41.6 [35.9, 47.3]
	gemini-2.5-flash	64.9 [63.9, 65.9]	69.4 [67.9, 71.0]	88.0 [87.0, 89.1]	76.1 [74.0, 78.0]	41.4 [40.5, 42.3]	68.0 [65.4, 70.5]	67.1 [66.5, 67.7]	44.1 [38.4, 50.0]
Others	kimi-k2.5	69.1 [68.0, 70.1]	68.9 [67.3, 70.5]	87.4 [86.2, 88.5]	73.5 [70.8, 76.0]	47.3 [46.1, 48.4]	71.9 [68.9, 74.7]	76.3 [75.6, 77.0]	58.3 [52.5, 63.8]
	glm-5.1	65.5 [64.4, 66.6]	57.9 [55.3, 60.4]	85.4 [84.1, 86.7]	75.8 [73.3, 78.2]	49.0 [47.8, 50.2]	69.5 [66.8, 72.0]	74.8 [74.0, 75.6]	46.3 [40.6, 52.0]
	doubao-seed-1.6	64.8 [63.8, 65.8]	66.0 [64.4, 67.6]	84.9 [83.7, 86.2]	74.7 [72.2, 77.1]	42.2 [41.2, 43.3]	61.5 [58.8, 64.2]	72.7 [72.0, 73.4]	51.4 [45.9, 57.0]
	qwen3-vl-32b-instruct	61.3 [60.3, 62.3]	60.9 [59.3, 62.5]	82.0 [80.7, 83.4]	73.5 [71.3, 75.5]	39.1 [38.4, 40.0]	59.4 [57.6, 61.3]	66.2 [65.5, 66.9]	48.0 [42.3, 53.8]

B Pairwise Paired-Bootstrap Significance on Overall Score

The point-estimate ranking in Table 3 is informative but does not by itself establish whether two adjacent models differ in a statistically meaningful sense. Table 6 reports the two-sided p-value for each ordered pair (i, j) of models under a paired bootstrap on the Overall score. For each of the $B = 10{,}000$ resamples we use the *same* per-task item indices to recompute every model's Overall score, so that the difference $D_b = \widehat{O}_i^{(b)} - \widehat{O}_j^{(b)}$ inherits the per-item correlation between the two models. The reported p-value is the standard two-sided percentile, $\hat{p} = 2 \cdot \min\big(\Pr(D_b \geq 0),\ \Pr(D_b \leq 0)\big)$, where the probabilities are taken over the bootstrap distribution. Bold cells flag $p < 0.05$; italic cells flag $0.05 \leq p < 0.10$; plain cells are non-significant. To keep the matrix readable, models are referenced by the integer IDs defined in Table 5; IDs follow the same provider-grouped ordering used throughout the paper.

Table 5: Model ID legend used in Table 6.

ID	Model	ID	Model
1	`gpt-5.4`	9	`kimi-k2.5`
2	`claude-4.6-opus`	10	`gpt-5-mini`
3	`claude-3.7-sonnet`	11	`gemini-3-flash`
4	`claude-4.6-sonnet`	12	`glm-5.1`
5	`gpt-5.2`	13	`gemini-2.5-flash`
6	`gemini-3.1-pro`	14	`doubao-seed-1.6`
7	`gpt-5`	15	`qwen3-vl-32b-instruct`
8	`gpt-5.1`	16	`gpt-4o`

The top of the ranking is a statistical tie. The most consequential pair for the headline claim is `gpt-5.4` (ID 1) versus `claude-4.6-opus` (ID 2): their paired-bootstrap p-value is 0.092, which fails to reject equality at the 0.05 level and only barely crosses the 0.10 threshold. `gpt-5.4` *is* significantly above the third-place `claude-3.7-sonnet` ($p_{1,3} = 0.020$) and `claude-4.6-sonnet` ($p_{1,4} = 0.010$), but its margin over `claude-4.6-opus` sits in the same regime as several within-cluster comparisons below. We therefore describe the leader as "statistically tied at the top" rather than as a clear winner in the main text.

Table 6: Pairwise paired-bootstrap two-sided p-values on the Overall benchmark score, $B = 10{,}000$ resamples with shared within-task item indices across models. **All entries are reported in units of 10^{-3}**, so a cell value of 92 means $p = 0.092$ and "< 1" means $p < 10^{-3}$. Rows and columns are indexed by the model IDs listed in Table 5. Bold: $p < 0.05$. Italic: $0.05 \leq p < 0.10$. Plain: $p \geq 0.10$. The matrix is symmetric ($p_{ij} = p_{ji}$); the diagonal is empty.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	—	*92*	**20**	**10**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
2	*92*	—	520	158	*55*	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
3	**20**	520	—	512	164	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
4	**10**	158	512	—	418	**1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
5	**<1**	*55*	164	418	—	**3**	**3**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
6	**<1**	**<1**	**<1**	**1**	**3**	—	331	**6**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
7	**<1**	**<1**	**<1**	**<1**	**3**	331	—	*85*	**3**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
8	**<1**	**<1**	**<1**	**<1**	**<1**	**6**	*85*	—	**47**	*68*	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
9	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**3**	**47**	—	989	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
10	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	*68*	989	—	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**
11	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	—	558	**34**	*78*	**<1**	**<1**
12	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	558	—	158	168	**<1**	**<1**
13	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**34**	158	—	852	**<1**	**<1**
14	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	*78*	168	852	—	**<1**	**<1**
15	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	—	**<1**
16	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	**<1**	—

A four-model cluster sits just below the leader. IDs 2–5 (`claude-4.6-opus`, `claude-3.7-sonnet`, `claude-4.6-sonnet`, `gpt-5.2`) are mutually indistinguishable: every pairwise comparison among them yields $p \in [0.158, 0.520]$, with the only borderline case being $p_{2,5} = 0.055$. The three Claude models in particular cannot be ordered statistically (all internal p-values ≥ 0.158). Combined with the leader pair above, this means the top five rows of Table 3 should be read as a *cluster* rather than as a strict ranking.

Lower-tier pairs that look close are also indistinguishable. Several mid-table neighbours fail to separate at conventional levels: $p_{6,7} = 0.331$ (`gemini-3.1-pro` vs. `gpt-5`), $p_{7,8} = 0.085$ (`gpt-5` vs. `gpt-5.1`), $p_{9,10} = 0.989$ (`kimi-k2.5` vs. `gpt-5-mini`), $p_{11,12} = 0.558$ (`gemini-3-flash` vs. `glm-5.1`), and $p_{13,14} = 0.852$ (`gemini-2.5-flash` vs. `doubao-seed-1.6`). These five clusters partition the leaderboard into five quasi-equivalence classes; differences within a class should not be treated as evidence of model superiority.

Cross-cluster gaps are robust. By contrast, every cross-cluster comparison—e.g. any leader-cluster model against any model below `gemini-3.1-pro`, or any mid-tier model against `qwen3-vl-32b-instruct` (ID 15) or `gpt-4o` (ID 16)—yields $p < 0.001$. The lowest two models are themselves strongly separated from each other ($p_{15,16} < 0.001$). Together these patterns indicate that AER-bench resolves coarse capability tiers unambiguously while leaving the finest within-tier ordering to be settled by additional data—a useful diagnostic for sample-size planning in future benchmark extensions.

C Judge Self-Preference Significance Analysis

The two LLM judges used for the open-generation tasks (`claude-4.6-sonnet` and `azure-gpt-5.2`) are themselves candidates in the benchmark. Reviewers reasonably worry that this creates a structural conflict of interest: each judge could systematically inflate the scores of models from its own family. The main text (Section 4.2) summarises the result; this appendix documents the full procedure, the per-model statistics, and the limitations.

Statistical setup. For every examinee m and every item i on a non-CI task, we observe two normalised scores in $[0, 1]$, one from each judge. Define the per-item gap

$$\Delta_{m,i} \;=\; s_{m,i}^{\text{claude-judge}} \;-\; s_{m,i}^{\text{gpt-judge}},$$

so that $\Delta > 0$ means the Claude judge is more lenient on item i than the GPT judge. Stacking across the six non-CI tasks gives $n = 1{,}469$ paired observations per examinee. We run two complementary tests:

1. **Per-model paired Wilcoxon signed-rank test** on $\{\Delta_{m,i}\}_{i=1}^{n}$ (two-sided), to detect whether *any* systematic offset exists between the two judges for examinee m.
2. **Family-level Mann–Whitney U test** on the per-model means $\overline{\Delta}_m$, comparing the three Anthropic candidates against the six OpenAI candidates. This is the test that directly probes *in-family* bias: if the Claude judge favoured Anthropic candidates over OpenAI candidates, Anthropic models would have systematically larger $\overline{\Delta}_m$ than OpenAI models.

Per-model results. Table 7 reports $\overline{\Delta}_m$ and the Wilcoxon p-value for each of the 16 examinees. With $n = 1{,}469$ paired items the Wilcoxon test is highly powered, so almost every model shows a statistically significant non-zero gap; the magnitude rather than the p-value is the quantity of interest.

Table 7: Per-examinee mean Claude-minus-GPT judge gap $\overline{\Delta}_m$ (in percentage points) with paired Wilcoxon signed-rank p-values (two-sided, $n = 1{,}469$ items per row). Positive values mean the Claude judge scores examinee m higher than the GPT judge does. The footnote reports family-level mean gaps and the Mann–Whitney U test comparing Anthropic vs. OpenAI candidates.

Examinee	Family	n	$\overline{\Delta}$ (pp)	Wilcoxon p
claude-3.7-sonnet	Anthropic	1469	+3.96	**<0.001**
claude-4.6-opus	Anthropic	1469	+2.41	**<0.001**
claude-4.6-sonnet	Anthropic	1469	+2.32	**<0.001**
gemini-2.5-flash	Google	1469	-1.43	**<0.001**
gemini-3-flash	Google	1469	+0.05	0.486
gemini-3.1-pro	Google	1469	-0.01	**<0.001**
gpt-4o	OpenAI	1469	-2.05	**<0.001**
gpt-5	OpenAI	1469	+1.69	**<0.001**
gpt-5-mini	OpenAI	1469	+1.60	**<0.001**
gpt-5.1	OpenAI	1469	+1.39	**<0.001**
gpt-5.2	OpenAI	1469	+2.00	**<0.001**
gpt-5.4	OpenAI	1469	+3.85	**<0.001**
doubao-seed-1.6	Other	1469	+1.56	**<0.001**
glm-5.1	Other	1469	+3.57	**<0.001**
kimi-k2.5	Other	1469	+3.53	**<0.001**
qwen3-vl-32b-instruct	Other	1469	-1.52	**<0.001**

Group means:
Anthropic $\bar{\Delta} = +2.90$ pp ($n_g = 3$);
OpenAI $\bar{\Delta} = +1.41$ pp ($n_g = 6$);
Other $\bar{\Delta} = +0.82$ pp ($n_g = 7$).
Mann–Whitney U test (Anthropic vs. OpenAI): $p = 0.095$.

A directional but underpowered in-family effect. The three Anthropic candidates receive an average Claude-judge bonus of $\overline{\Delta} = +2.90$ pp, roughly twice the OpenAI candidates' $+1.41$ pp and more than three times the "Other" families' $+0.82$ pp. The direction of the effect is exactly what the self-preference hypothesis predicts. However, the Mann–Whitney U test on the per-model means yields $p = 0.095$, falling short of the conventional 0.05 threshold: with only $n_g = 3$ Anthropic and $n_g = 6$ OpenAI candidates the family-level test has limited power. We therefore *cannot* reject the null of no in-family bias, but we equally *cannot* claim that the apparent ~ 1.5 pp Anthropic–OpenAI gap is zero.

Why the effect is small in practice. Two facts bound the practical impact of this directional bias on the benchmark rankings. First, the quantity that matters for the leaderboard is the *between-family* component (~ 1.5 pp), not the global Claude-vs-GPT lenience offset ($+1.43$ pp averaged across all 16 candidates), because the latter is absorbed by averaging the two judges rather than picking either one. Second, the bias is far smaller than the leaderboard's gross dynamic range: the top-to-bottom Overall gap is roughly 21 pp (`gpt-5.4` 75.0% vs. `gpt-4o` 54.0%), and the only Anthropic–OpenAI comparisons whose ordering a ~ 1.5 pp shift could plausibly disturb already sit inside the within-cluster ties that the pairwise paired-bootstrap analysis (Appendix B) flags as statistically indistinguishable (most prominently `gpt-5.4` vs. `claude-4.6-opus`, $p = 0.092$). All Anthropic-vs-OpenAI pairs that *are* significant in Appendix B have point-estimate gaps that substantially exceed the residual family-level bias, so no cross-tier ranking claim in the main text is at risk of flipping under a 1.5 pp adjustment.

What this analysis does not establish. The two-judge protocol attenuates but does not eliminate the conflict-of-interest concern: the test above is necessary but not sufficient. A truly independent third-party judge—e.g. `gemini-3.1-pro`, which is a candidate but not currently a judge—would be needed to fully decouple judging from candidate selection. We treat the addition of a third independent judge as the single highest-value follow-up experiment for future versions of AER-bench, and discuss it explicitly in the limitations section.

D Robustness Controls for the Historical Concept-Drift Finding

The main-text Finding 3 (Section 4) reframes our Concept Identification (CI) results as a *model-stratified, length-conditional* "modernity bias" rather than a uniform property of contemporary LLMs. This appendix documents the two robustness controls behind the reframing: an item-clustered OLS regression with chunk-length control, and a per-decade $\times$ per-model breakdown of CI accuracy with bootstrap confidence intervals.

Why these controls are needed. The aggregate $\text{Era}_3 - \text{Era}_2$ gap of $\sim +5$ pp in Figure 6 can be confounded by at least three non-conceptual factors: (i) systematic length differences between eras; (ii) a 30-year era partition that may hide decade-level heterogeneity; and (iii) aggregation across 16 models with potentially very different behaviours. We re-analyse the CI data along all three axes.

Data and setup. We recover each CI item's publication year and source-chunk text from the benchmark's raw CI files, use chunk character count as a length proxy, and bin items into five decades. Of 160 CI items, 146 carry a usable year + chunk pair; the decade breakdown is $\{10, 22, 27, 43, 44\}$ items. Per-model bootstrap intervals on each decade use $B = 10{,}000$ item-level resamples.

OLS with chunk-length control. We estimate the panel regression

$$\text{score}_{m,i} \;=\; \alpha_m + \beta_{\text{Era}_3} \mathbb{1}[\text{Era}_3]_i + \beta_{\log \text{len}} \log(\text{len})_i + \beta_{\times} \mathbb{1}[\text{Era}_3]_i \cdot \log(\text{len})_i + \varepsilon_{m,i},$$

with model fixed effects α_m, mean-centred $\log(\text{len})_i$, and item-clustered standard errors. We obtain $\hat{\beta}_{\text{Era}_3} = +0.0275$ ($p = 0.442$), $\hat{\beta}_{\log \text{len}} = -0.1502$ ($p < 10^{-5}$), $\hat{\beta}_{\times} = +0.0572$ ($p = 0.319$). Two facts follow: (a) the unconditional $+5$ pp era gap shrinks to $+2.75$ pp and is no longer distinguishable from zero once length is controlled for; (b) longer paragraphs are uniformly harder (~ -15 pp per unit of log-length), the dominant CI difficulty signal. The $\text{Era}_3 \times$ length interaction is positive but not individually significant.

Table 8: Length-stratified CI accuracy by era. "Gap" is the per-quartile $\text{Era}_3 - \text{Era}_2$ difference in mean accuracy; the bottom row reports the OLS coefficients. After controlling for length, the Era_3 main effect is no longer distinguishable from zero.

Length quartile	Era2 acc. (%)	Era3 acc. (%)	Gap (pp)	n_{Era2}	n_{Era3}
Q1 (shortest)	69.1	67.3	-1.8	304	288
Q2	49.4	47.3	-2.1	240	336
Q3	48.3	57.5	+9.2	256	320
Q4 (longest)	42.3	55.0	+12.7	144	448

OLS with model fixed effects and item-clustered standard errors:
$\beta_{\text{Era3}} = +0.0275$ ($p = 0.442$);
$\beta_{\log\ len} = -0.1502$ ($p = 9.16 \times 10^{-6}$);
$\beta_{\text{Era3} \times \log\ len} = +0.0572$ ($p = 0.319$).

Length-quartile stratification. Table 8 reveals a non-linear pattern the OLS coefficients hide: the per-quartile $\text{Era}_3 - \text{Era}_2$ gap is -1.8 pp on Q1, -2.1 pp on Q2, $+9.2$ pp on Q3, and $+12.7$ pp on Q4. The two shortest quartiles *reverse* the sign of the modernity bias, while the two longest account for essentially all of the aggregate $+5$ pp gap—consistent with representational gaps that surface only when models must reason over several inter-related sentences of pre-1930 prose.

Per-decade, per-model heterogeneity. Table 9 shows that aggregating across models obscures the dominant regularity. Along the 1900–1909 vs. 1940–1950 axis, models split into two clean groups:

- **Earliest-decade-strong** (`claude-4.6-sonnet`, `claude-4.6-opus`, `claude-3.7-sonnet`, `gemini-3.1-pro`, `gpt-5.2`, `kimi-k2.5`) score *higher* on 1900–1909 than on 1940–1950, with old-to-recent drops between -3 and -22 pp (extreme: `claude-4.6-sonnet` 86.2% vs. 64.2%).
- **Modern-skewed** (`gpt-5.4`, `gpt-5`, `gpt-5-mini`, `gemini-3-flash`, `gemini-2.5-flash`, `glm-5.1`) show the canonical pattern: 19–41% on 1900–1909 rising roughly monotonically to 46–59% on 1940–1950.

Table 9: Per-decade CI accuracy (%) with item-level 95% bootstrap confidence intervals ($B = 10{,}000$). Models are ordered by their 1900–1909 score; the *Items* row gives the per-decade sample size, so intervals on the $n = 10$ column are correspondingly wide.

Model	1900-09	1910-19	1920-29	1930-39	1940-50
Items	10	22	27	43	44
claude-4.6-sonnet	86.2 [65.0, 100.0]	80.1 [66.5, 92.6]	78.2 [66.7, 88.4]	67.2 [57.6, 76.2]	64.2 [52.8, 75.6]
claude-4.6-opus	82.5 [62.5, 97.5]	83.5 [71.0, 94.3]	69.4 [54.6, 82.9]	69.5 [59.9, 78.5]	65.3 [54.0, 75.9]
claude-3.7-sonnet	76.2 [48.8, 97.5]	80.7 [66.5, 92.6]	79.2 [67.1, 89.4]	64.8 [54.4, 75.0]	67.3 [55.7, 77.6]
gemini-3.1-pro	81.2 [58.8, 96.2]	75.6 [59.1, 90.3]	75.0 [63.4, 86.1]	63.7 [54.1, 72.7]	72.2 [63.6, 80.1]
gpt-5.2	61.3 [32.5, 86.2]	66.5 [47.7, 83.5]	68.1 [53.2, 81.9]	57.0 [46.8, 67.2]	54.8 [44.0, 65.9]
kimi-k2.5	61.3 [33.8, 85.0]	51.7 [33.5, 69.9]	61.1 [46.8, 74.5]	58.4 [49.1, 67.4]	57.4 [47.4, 67.3]
gpt-5.4	41.2 [15.0, 68.8]	52.8 [35.2, 71.0]	56.9 [41.7, 71.8]	64.8 [55.8, 73.8]	59.4 [48.0, 70.5]
doubao-seed-1.6	66.2 [41.2, 87.5]	45.5 [27.8, 63.6]	54.2 [39.8, 68.5]	45.6 [36.3, 54.9]	51.7 [41.8, 61.9]
gpt-5.1	65.0 [40.0, 86.2]	51.1 [33.0, 69.3]	42.6 [28.2, 57.4]	52.3 [42.7, 61.9]	49.7 [38.9, 60.5]
gpt-4o	50.0 [22.5, 76.2]	44.9 [27.3, 63.6]	51.9 [37.0, 66.7]	57.3 [47.4, 66.6]	52.6 [41.8, 62.8]
qwen3-vl-32b-instruct	46.2 [20.0, 72.5]	42.6 [25.0, 60.8]	51.9 [37.0, 66.2]	50.0 [40.4, 59.6]	45.2 [35.2, 55.7]
gpt-5	31.2 [7.5, 57.5]	31.8 [15.9, 49.4]	45.8 [31.5, 60.6]	52.9 [42.1, 63.1]	56.2 [45.2, 67.6]
glm-5.1	31.2 [6.2, 57.5]	31.8 [15.9, 48.9]	45.4 [31.0, 59.7]	48.0 [37.8, 57.8]	55.7 [45.7, 65.9]
gemini-2.5-flash	30.0 [6.2, 56.2]	29.5 [14.2, 46.6]	43.1 [28.7, 57.4]	48.8 [38.1, 58.7]	52.8 [42.3, 63.4]
gpt-5-mini	18.8 [0.0, 38.8]	29.0 [14.2, 45.5]	47.2 [32.9, 62.0]	53.2 [43.3, 63.1]	46.9 [35.8, 57.7]
gemini-3-flash	18.8 [0.0, 37.5]	32.4 [15.9, 50.0]	40.3 [25.9, 54.2]	49.4 [38.7, 59.9]	46.3 [34.9, 57.1]

Mid-tier models (`doubao-seed-1.6`, `gpt-5.1`, `qwen3-vl-32b-instruct`, `gpt-4o`) sit between the groups with comparatively flat profiles. Earliest-decade intervals are wide ($n = 10$), so the $\sim$20 pp within-row swings should be read alongside the 1–2 pp swings on the densely-sampled 1940–1950 column.

What the reframing buys us. Together, the three controls support a much narrower claim than "current AI systems exhibit historical concept drift": the era main effect vanishes after length control ($p = 0.442$); the descriptive gap concentrates in the longest two quartiles; and the per-decade structure differs sharply by model, with frontier Anthropic / Gemini-3.1-pro / GPT-5.2 systems scoring *better* on the earliest decade than on the most recent. We therefore frame the CI task as a *diagnostic for which models still carry a representational blind spot on early-20th-century economic prose*, rather than as evidence of a uniform property of contemporary LLMs. The same controls flag the two most informative follow-ups: (i) inter-annotator agreement on the 1900–1909 subset, where bootstrap intervals are widest, and (ii) controlled experiments holding chunk length fixed across eras.

E Detailed Rubric Definitions

This appendix provides detailed rubric definitions for all LLM-as-a-Judge tasks in AER-bench. Each task employs multi-dimensional scoring to prevent holistic biases and ensure fine-grained evaluation of model capabilities.

E.1 Mathematical Modeling (MM)

Mathematical Modeling evaluates the ability to derive equations, solve optimization problems, and perform formal mathematical reasoning in economic contexts. Each question is scored across four dimensions (0–10 points each, total 40 points):

Correctness (Weight: 4) Assesses whether formulas are mathematically equivalent to the reference answer and derivation steps are valid.

- **9–10**: All formulas are mathematically equivalent to the reference; derivation steps are valid.
- **6–8**: Minor errors (sign, index, simplification) but core structure is correct.
- **3–5**: Significant errors but partial understanding of the approach is evident.
- **0–2**: Fundamentally incorrect or mostly missing.

Notation Consistency (Weight: 2) Evaluates whether symbols match the problem setup and are used consistently throughout the derivation.

- **9–10**: All symbols match the problem setup and are used consistently.
- **6–8**: Minor inconsistencies that do not affect readability.
- **3–5**: Some symbols undefined or used inconsistently.
- **0–2**: Major notation confusion or deviation from given symbols.

Completeness (Weight: 3) Checks whether all intermediate steps are shown and a reader can verify each step.

- **9–10**: All intermediate steps shown; a reader can verify each step.
- **6–8**: Key steps present but some intermediate algebra omitted.
- **3–5**: Major steps missing; jumps to conclusions without justification.
- **0–2**: Only final answer given, or derivation is fragmentary.

Economic Intuition (Weight: 1) Assesses whether the candidate provides clear interpretation of results in economic terms and explains what the mathematics means.

- **9–10**: Clear interpretation of results in economic terms; explains what the math means.
- **6–8**: Some economic interpretation provided.
- **3–5**: Minimal economic context; mostly mechanical derivation.
- **0–2**: No economic interpretation at all.

E.2 Paper Abstraction (PA)

Paper Abstraction evaluates the ability to synthesize a condensed paper body (approximately 30% of original length, excluding the abstract) into a coherent abstract that captures key contributions, methods, and findings. Scoring is based on four dimensions (0–10 points each, total 40 points):

Coverage (Weight: 4) Measures whether all or nearly all reference key points (atomic semantic propositions extracted from the original abstract) are mentioned or clearly implied.

- **9–10**: All or nearly all reference key points are mentioned or clearly implied.
- **6–8**: Most key points covered; 1–2 minor omissions.

- **3–5**: Several important points missing.
- **0–2**: Only 1–2 points covered, or major elements absent.

Accuracy (Weight: 4) Assesses factual correctness of all stated claims per the paper content.

- **9–10**: All stated claims are factually correct per the paper content.
- **6–8**: Minor inaccuracies that do not fundamentally mislead.
- **3–5**: Some factual errors or misrepresentations of results.
- **0–2**: Major factual errors that distort the paper's findings.

Conciseness (Weight: 3) Evaluates whether the abstract stays within the specified word limit with no redundancy.

- **9–10**: Within word limit; no redundancy; every sentence adds value.
- **6–8**: Mostly concise; minor redundancy or slightly over limit.
- **3–5**: Noticeably wordy or exceeds word limit significantly.
- **0–2**: Excessively long or highly redundant.

Academic Style (Weight: 1) Judges whether the abstract reads like a published abstract in a top economics journal.

- **9–10**: Reads like a published abstract in a top economics journal.
- **6–8**: Generally academic but with some informal phrasing.
- **3–5**: Mixed register; some casual language or poor structure.
- **0–2**: Not academic in style.

E.3 Literature Review (LR)

Literature Review evaluates the ability to reconstruct an introduction section that aligns with the paper's citation network and argumentation structure. This task employs a hybrid scoring approach combining citation recall and writing quality.

Citation Score (Weight: 3) Measures weighted recall of citations across three tiers:

- **Tier 1 (weight 0.8)**: Core citations indispensable to the argument, typically from top journals or highly influential sources.
- **Tier 2 (weight 0.2)**: Supplementary citations from reputable academic sources.
- **Tier 3 (weight 0.0)**: Optional citations with lower relevance or authority; not scored.

The citation score is computed as: citation_score $= 0.8 \times$ Tier1_recall $+ 0.2 \times$ Tier2_recall.

Rubric Score (Weight: 2) Assesses writing quality across paper-specific dimensions tailored to the introduction's argumentation structure. Common dimensions include:

- **Research question clarity**: Is the core question clearly stated?
- **Literature positioning**: Does the introduction position the paper within existing literature?
- **Contribution statement**: Are the paper's contributions explicitly articulated?
- **Methodological preview**: Is the research approach previewed appropriately?
- **Logical flow**: Does the introduction follow a coherent argumentative structure?

Each dimension is scored 0–10, with weights summing to 100. The rubric is dynamically constructed for each paper based on its specific introduction structure.

E.4 Model Generation (MG)

Model Generation evaluates the ability to construct a complete formal economic model from a desymbolized narrative description. Unlike other tasks, MG uses a flexible element-based rubric where the number and type of required elements vary by question.

Element Rubric (Flexible Dimensions) Each model is decomposed into constituent elements (agents, objective functions, constraints, equilibrium concepts, first-order conditions). Each element is scored based on presence and correctness:

- **Full (100%)**: Element is present, correctly formalized, uses proper mathematical notation, and is economically meaningful.
- **Substantial (70%)**: Element is present and mostly correct, but has minor issues (e.g., missing a variable, slight notation inconsistency, or incomplete specification).
- **Partial (40%)**: Element is mentioned or attempted, but has significant issues (e.g., wrong functional form, missing key variables, or conceptually flawed formalization).
- **Minimal (15%)**: Element is only vaguely referenced without proper formalization.
- **Missing (0%)**: Element is not present at all.

The total score is the sum of all element scores, normalized to 0–100. The maximum possible score and element weights are determined by the reference model's complexity. For example, a model with 2 agents, 2 objective functions, 3 constraints, 1 equilibrium concept, and 2 FOCs would have 10 elements, each contributing proportionally to the total score.

E.5 Economic Analysis (EA)

Economic Analysis evaluates the ability to interpret empirical or theoretical results with substantive economic reasoning. Given a purely factual data description (e.g., regression coefficients, significance levels, trends), the model must produce a professional analysis paragraph. Scoring is based on six dimensions (0–10 points each), with applicability determined per question:

Accuracy (Weight: 2) Assesses whether numerical descriptions and trend characterizations are factually correct.

- **9–10**: All numerical and directional claims are accurate; no fabrication.
- **6–8**: Minor inaccuracies that do not fundamentally mislead.
- **3–5**: Some factual errors in data description.
- **0–2**: Major errors or fabricated numbers.

Economic Interpretation (Weight: 2) Evaluates whether the analysis provides substantive economic meaning (e.g., magnitude interpretation, heterogeneity discussion).

- **9–10**: Rich interpretation of economic significance and practical implications.
- **6–8**: Reasonable interpretation provided.
- **3–5**: Superficial or generic interpretation.
- **0–2**: No substantive interpretation.

Mechanism (Weight: 2) Assesses whether economic channels or causal pathways are identified.

- **9–10**: Clear identification of driving mechanisms.
- **6–8**: Some mechanisms discussed but not fully developed.
- **3–5**: Vague or incomplete mechanism discussion.
- **0–2**: No mechanism identified.

Theory Connection (Weight: 1.5) Evaluates whether results are connected to theoretical predictions or prior work.

- **9–10**: Strong theoretical grounding with specific references.
- **6–8**: Some theoretical context provided.
- **3–5**: Weak or generic theory connection.
- **0–2**: No theory connection.

Policy Implication (Weight: 1.5) Assesses whether policy implications are discussed appropriately (only applicable when results support policy discussion).

- **9–10**: Nuanced, well-grounded policy discussion.
- **6–8**: Reasonable policy points raised.
- **3–5**: Superficial policy discussion.
- **0–2**: No policy discussion.
- **N/A**: Not applicable for this result.

Writing Quality (Weight: 1) Judges whether the analysis is well-structured and clearly written.

- **9–10**: Publication-ready prose; logical flow; precise language.
- **6–8**: Clear and organized; minor style issues.
- **3–5**: Readable but poorly organized or imprecise.
- **0–2**: Disorganized or unclear.

Note: Not all dimensions are applicable to every question. For example, a purely theoretical result may not have policy implications. The normalized score is computed as: $\text{normalized\_score} = \frac{\text{total}}{\text{max\_possible}}$, where max_possible counts only applicable dimensions (each worth 10 points).

E.6 Judge Model Configuration

All LLM-as-a-Judge evaluations in AER-bench use the following judge models:

- **claude-4.6-sonnet** (Anthropic)
- **gpt-5.2** (OpenAI, accessed via Azure)

For each generative task, both judges independently score all dimensions. The final item score is the arithmetic mean of the two judges' scores. Section 4.2 presents a systematic comparison of inter-judge agreement, demonstrating a mean absolute difference of 2.6 percentage points across all models and tasks, validating the robustness of this multi-judge approach.

F Detailed Performance Analysis by Task Dimension

This appendix provides a comprehensive breakdown of model performance across all seven task dimensions in AER-bench. Figure 7 presents a visual summary of the overall ranking table (Table 3), showing task-level performance for all evaluated models grouped by provider. Each subsection then presents a heatmap visualization showing dimension-level scores for all evaluated models, followed by detailed analysis of performance patterns, strengths, and weaknesses.

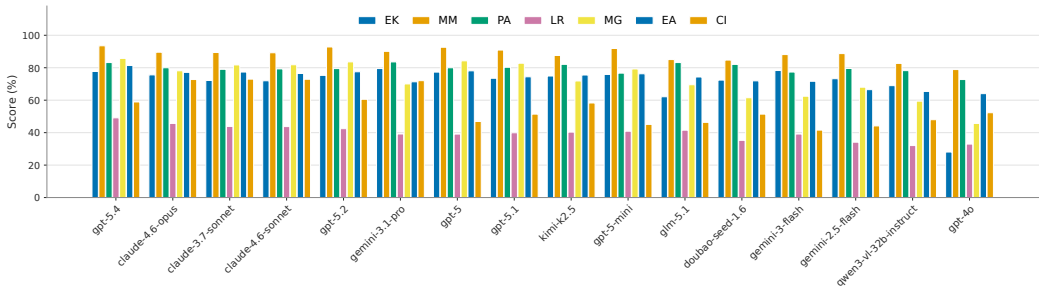

Figure 7: Task-level performance across all evaluated models, grouped by provider. This figure provides a visual complement to Table 3, showing the distribution of scores across the seven task dimensions (EK, MM, PA, LR, MG, EA, CI) for each model.

F.1 Economic Knowledge (EK)

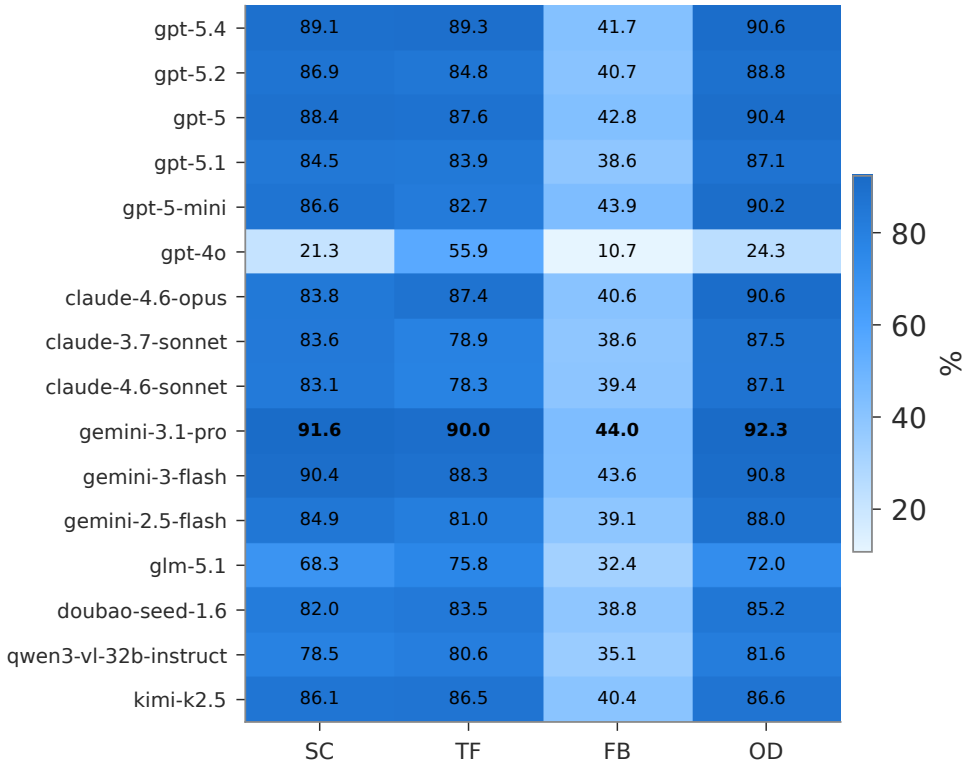

Figure 8: Performance breakdown across Economic Knowledge (EK) question types: Single Choice (SC), True/False (TF), Fill-in-Blank (FB), and Ordering (OD). Values represent percentage scores (0–100).

The Economic Knowledge task evaluates models across four question formats: single choice, true/false, fill-in-blank, and ordering. As shown in Figure 8, performance varies significantly across these formats. Google's `gemini-3.1-pro` achieves the highest overall EK score (79.5%), demonstrating particularly strong performance on single-choice questions. OpenAI's GPT-5 series models show consistent performance across all four formats, with `gpt-5.4` and `gpt-5` both scoring above 77%.

Notably, fill-in-blank questions prove most challenging across all models, with average scores approximately 40–50 percentage points lower than single-choice questions (e.g., `gemini-3.1-pro` scores 91.6% on single choice but only 44.0% on fill-in-blank). This suggests that open-ended knowledge recall is fundamentally more difficult than recognition-based tasks. Ordering questions generally yield the highest scores across models (up to 92.3%), likely because they test relational reasoning among familiar economic concepts. True/false questions also show high performance, though slightly below ordering.

Among the lower-performing models, `gpt-4o` shows a marked weakness in this dimension (28.1% overall), particularly struggling with fill-in-blank and ordering tasks. This indicates that older model generations lack the economic domain knowledge depth required for AER-bench tasks.

F.2 Mathematical Modeling (MM)

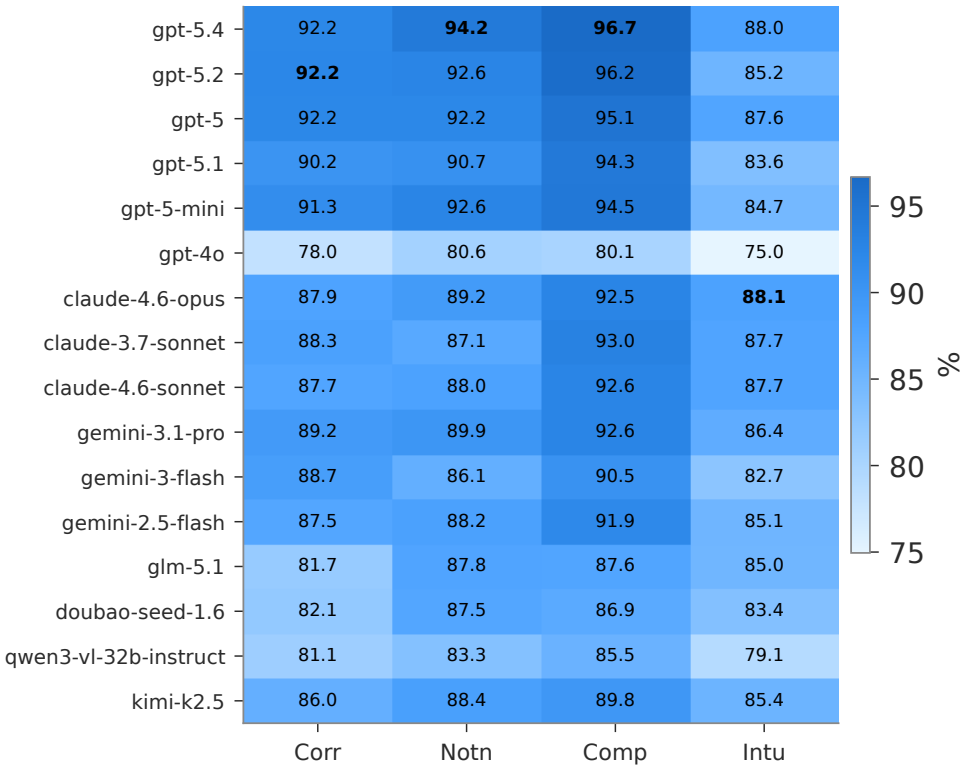

Figure 9: Performance breakdown across Mathematical Modeling (MM) dimensions: Correctness (Corr), Notation (Notn), Completeness (Comp), and Economic Intuition (Intu). Values represent percentage scores (0–100).

Mathematical Modeling represents the highest-scoring task across all models, with most frontier models exceeding 85% overall performance. Figure 9 reveals that `gpt-5.4` achieves exceptional performance (93.5%), with near-perfect scores across all four evaluation dimensions: correctness, notation, completeness, and economic intuition.

The heatmap shows remarkably consistent performance across dimensions for top-tier models. Correctness scores are universally high (>90%) for GPT-5 series models (`gpt-5.4`, `gpt-5.2`, `gpt-5`

all at 92.2%, `gpt-5-mini` at 91.3%, `gpt-5.1` at 90.2%), indicating strong mathematical reasoning capabilities. Claude 4.6 models achieve slightly lower but still strong correctness scores (87.7–87.9%). Notation scores, which assess proper use of mathematical symbols and formatting, are similarly high, suggesting models have learned standard economic notation conventions from training data.

Completeness and economic intuition dimensions show slightly more variation. Economic intuition, which evaluates whether models provide meaningful economic interpretation alongside mathematical derivations, proves most challenging. Models like `gemini-3-flash` and `qwen3-vl-32b-instruct` show a 5–8 point gap between correctness and intuition scores, indicating they can solve equations correctly but struggle to explain economic meaning.

The uniformly high MM scores across models suggest that mathematical manipulation and symbolic reasoning are well-developed capabilities in current LLMs, likely benefiting from extensive training on mathematical and scientific texts.

F.3 Paper Abstraction (PA)

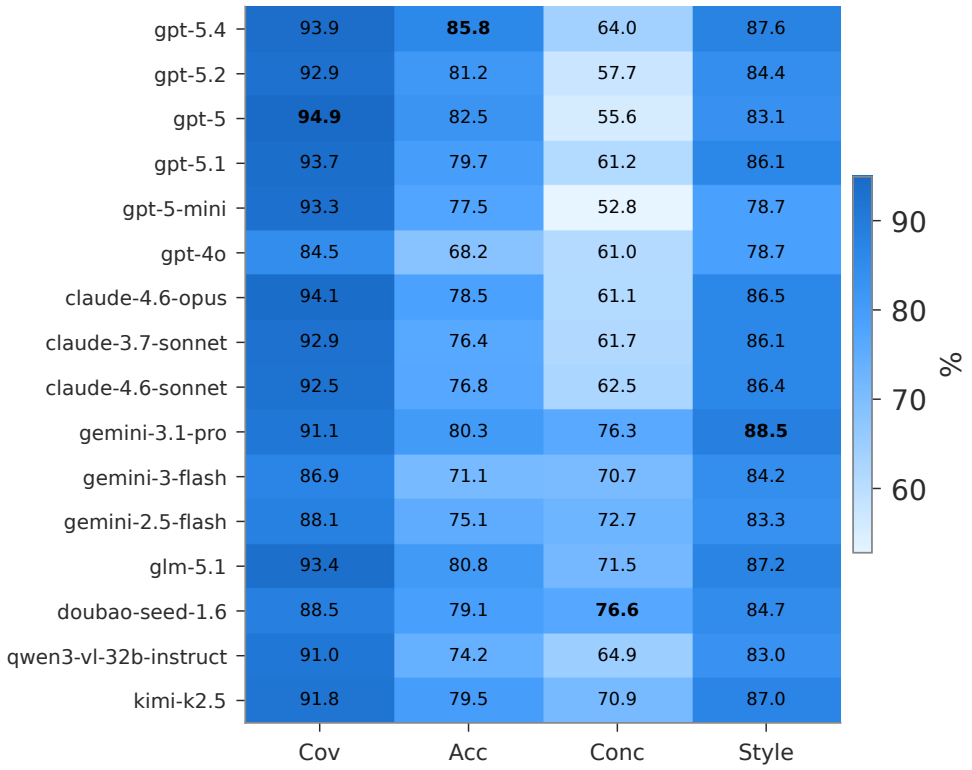

Figure 10: Performance breakdown across Paper Abstraction (PA) dimensions: Coverage (Cov), Accuracy (Acc), Conciseness (Conc), and Academic Style (Style). Values represent percentage scores (0–100).

Paper Abstraction tasks require models to summarize and analyze economic research papers across four dimensions: coverage, accuracy, conciseness, and academic style. Figure 10 shows that `gemini-3.1-pro` leads this task (83.6%), with particularly strong performance in coverage and accuracy dimensions.

Coverage scores, measuring whether summaries capture all key contributions, show the widest performance spread among dimensions. Top models (`gpt-5`, `gpt-5.4`, `claude-4.6-opus`, `gpt-5.1`, `glm-5.1`) achieve 91–95% coverage, while mid-tier models range from 86–91%. This suggests that identifying salient information from lengthy academic papers requires sophisticated comprehension capabilities.

Accuracy scores are generally high across all models, indicating that when models do extract information, they rarely introduce factual errors. However, conciseness proves challenging—many models produce verbose summaries that include excessive detail. The heatmap reveals that Claude models (`claude-4.6-opus`, `claude-3.7-sonnet`, `claude-4.6-sonnet`) show particularly balanced performance across all four dimensions, suggesting well-calibrated summarization capabilities.

Academic style scores, evaluating whether summaries use appropriate scholarly language and structure, correlate strongly with overall model capability. Frontier models consistently score above 80%, while smaller or older models (`gpt-4o`, `qwen3-vl-32b-instruct`) drop below 70%, indicating that academic writing conventions require substantial model capacity to master.

F.4 Literature Review (LR)

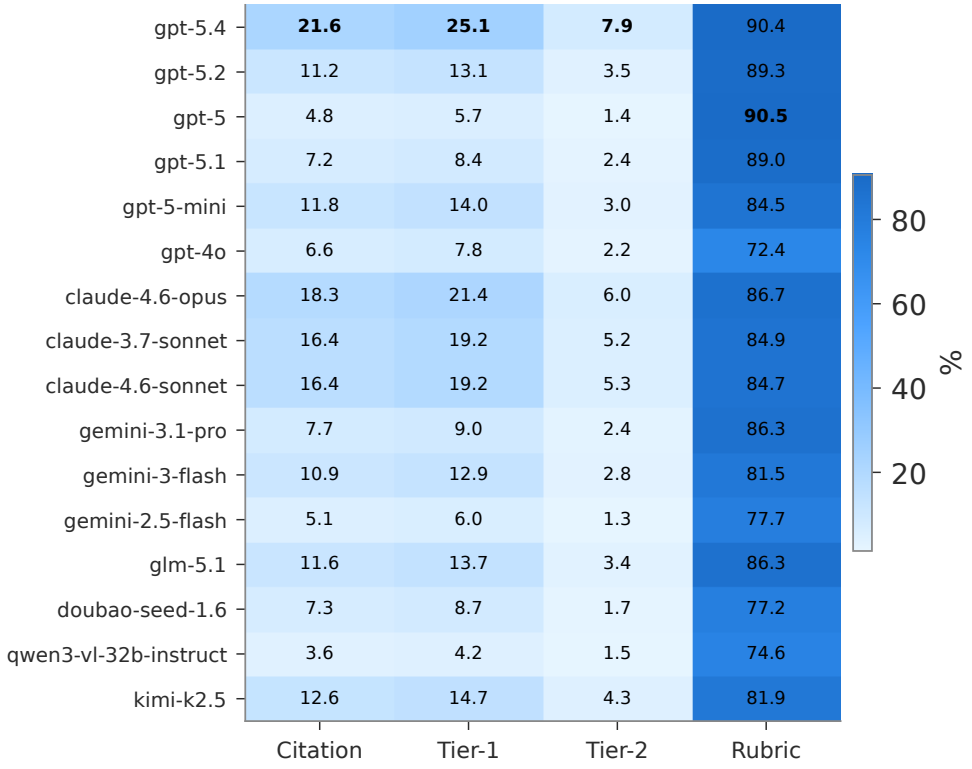

Figure 11: Performance breakdown across Literature Review (LR) dimensions: Citation Score (Cite) and Rubric Score (Rubric). Values represent percentage scores (0–100).

Literature Review is the most challenging task in AER-bench, with the highest-performing model (`gpt-5.4`) achieving only 49.1%. Figure 11 decomposes performance into citation score (accuracy of cited papers) and rubric score (quality of synthesis and analysis). The heatmap reveals that both dimensions are extremely difficult, with most models scoring below 45% on each.

Citation scores measure whether models correctly identify and cite relevant papers from a provided reference list. The extremely low scores indicate that models struggle with precise bibliographic matching and relevance assessment. Even the best-performing model, `gpt-5.4`, achieves only 21.6% citation accuracy, followed by `claude-4.6-opus` at 18.3%. Most models score below 15%, suggesting that fine-grained citation retrieval and bibliographic matching remain a fundamental weakness for current LLMs.

Rubric scores, evaluating the quality of literature synthesis, thematic organization, and critical analysis, are substantially higher. Top models achieve 85–90% rubric scores (e.g., `gpt-5` at 90.5%, `gpt-5.4` at 90.4%), indicating that models can produce well-structured literature review prose when the task does not require precise citation matching. The stark contrast between citation and rubric

scores reveals that the primary bottleneck in literature review is not analytical writing quality but rather the ability to accurately identify and attribute relevant sources.

Interestingly, the performance gap between frontier and mid-tier models is smaller in LR than in other tasks. This suggests that literature review capabilities scale less predictably with model size and training, possibly because the task requires specialized academic reasoning patterns underrepresented in general training data.

F.5 Method Generation (MG)

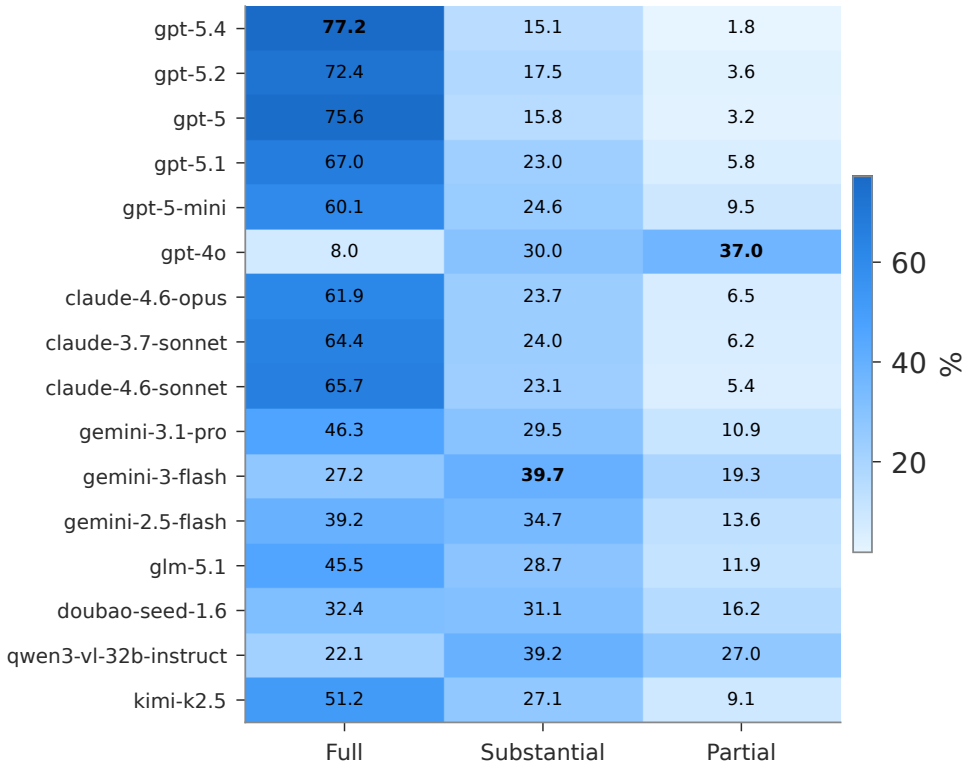

Figure 12: Performance breakdown across Method Generation (MG) rubric elements. Due to the flexible rubric structure, dimension labels vary by specific task item. Values represent percentage scores (0–100).

Method Generation tasks evaluate models' ability to design novel research methodologies, including identification strategies, data collection plans, and empirical specifications. Figure 12 shows performance across multiple rubric elements, which vary by specific task item due to the flexible evaluation structure.

`gpt-5.4` leads this task (85.8%), followed closely by `gpt-5` (84.2%) and `gpt-5.1` (82.8%). The heatmap reveals that top models achieve consistently high scores (80–90%) across most rubric elements, indicating strong methodological reasoning. However, certain elements—particularly those requiring creative problem-solving or domain-specific knowledge of econometric techniques—show greater performance variation.

Claude models demonstrate balanced performance across rubric elements, with `claude-3.7-sonnet` achieving 81.7%. Gemini models show more uneven performance, with `gemini-3.1-pro` scoring only 70.0% overall despite strong performance in other tasks. This suggests that method generation requires different capabilities than comprehension or analysis tasks.

Lower-tier models (`doubao-seed-1.6`, `qwen3-vl-32b-instruct`) score below 65%, indicating that generating methodologically sound research designs requires substantial reasoning capacity.

The task demands understanding of causal inference principles, data requirements, and econometric techniques—knowledge that appears concentrated in frontier models.

F.6 Economic Analysis (EA)

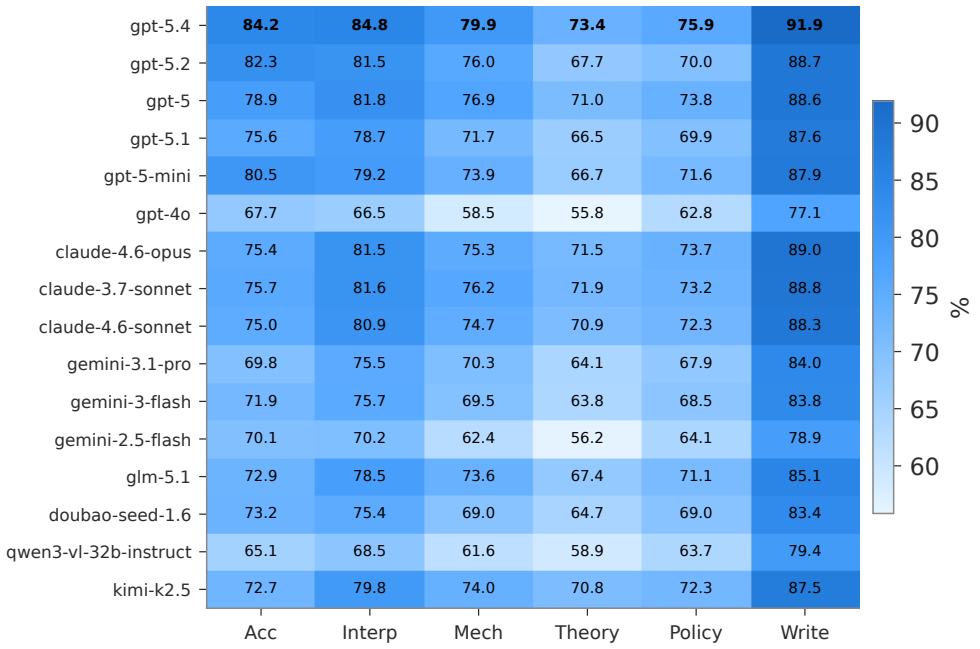

Figure 13: Performance breakdown across Economic Analysis (EA) dimensions: Accuracy (Acc), Economic Interpretation (Interp), Mechanism (Mech), Theory Connection (Theory), Policy Implication (Policy), and Writing Quality (Write). Values represent percentage scores (0–100).

Economic Analysis tasks require models to interpret regression results, explain economic mechanisms, connect findings to theory, and discuss policy implications. Figure 13 shows performance across six dimensions, revealing substantial variation in model capabilities.

`gpt-5.4` achieves the highest overall EA score (81.4%), with particularly strong performance in accuracy and writing quality dimensions. The heatmap shows that accuracy scores—measuring correct interpretation of statistical results—range from 65–84% across models, with frontier models (`gpt-5.4` at 84.2%, `gpt-5.2` at 82.3%, `gpt-5-mini` at 80.5%) scoring above 80%, indicating solid quantitative reasoning. However, deeper analytical dimensions show more variation.

Economic interpretation and mechanism explanation prove moderately challenging, with most models scoring 70–85%. These dimensions require understanding causal pathways and economic logic beyond statistical patterns. Theory connection is the most challenging EA dimension, with scores ranging from 56–73%. Even top models like `gpt-5.4` achieve only 73.4% on theory connection, suggesting that linking empirical findings to established theoretical frameworks requires sophisticated domain knowledge.

Policy implication scores are slightly higher than theory connection (ranging 63–76%), but still represent a significant challenge. This dimension requires extrapolating from research findings to real-world applications, considering practical constraints and normative judgments.

Writing quality scores are uniformly high (>75%) across frontier models, indicating that models can produce well-structured, coherent analytical text. The strong correlation between writing quality and overall EA performance suggests that clear communication and analytical depth are closely linked capabilities.

F.7 Concept Identification (CI)

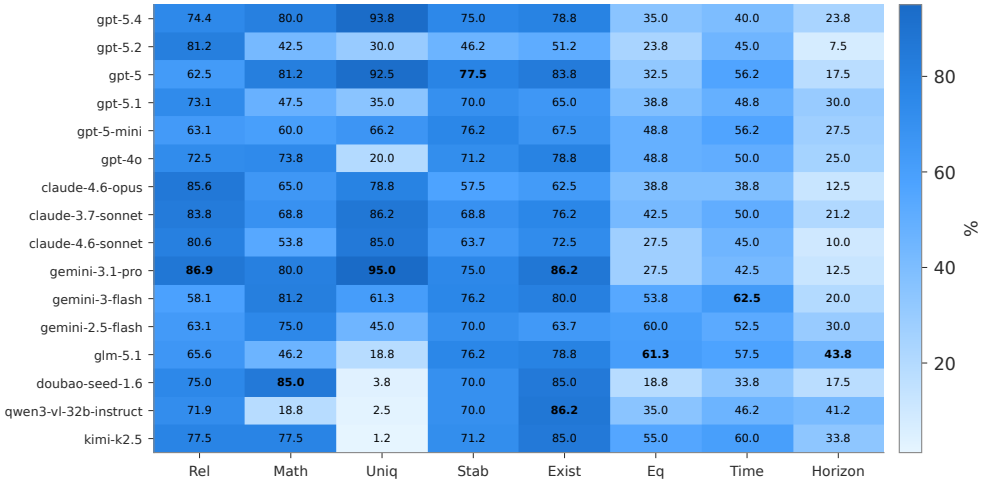

Figure 14: Performance breakdown across Concept Identification (CI) dimensions: Relevant, Math Formalization, Uniqueness, Stability, Existence, Equilibrium Scene, Time Characteristic, and Time Horizon. Values represent percentage scores (0–100).

Concept Identification tasks evaluate models' ability to identify and classify economic concepts from historical AER papers, assessing understanding across eight dimensions. Figure 14 reveals that this task shows the highest performance variance across models, with `claude-3.7-sonnet` leading (73.0%) while `gpt-5` scores only 46.9%.

The heatmap shows that Claude models (`claude-4.6-opus`, `claude-3.7-sonnet`, `claude-4.6-sonnet`) consistently outperform other providers on CI tasks, achieving 72.7–73.0% overall scores. However, their performance varies dramatically across individual dimensions, ranging from as low as 10% (time horizon for `claude-4.6-sonnet`) to as high as 86.2% (uniqueness for `claude-3.7-sonnet`). This suggests that Anthropic's training or architecture may be particularly well-suited for certain aspects of conceptual classification and historical document understanding, while struggling with temporal reasoning dimensions.

Relevance scores are generally the highest across models (ranging 58–87%), indicating that identifying whether a concept is economically relevant is relatively straightforward. However, other dimensions show much greater variance and difficulty. Math formalization scores range from 18.8% to 85.0%, with some models excelling (`doubao-seed-1.6`: 85.0%, `gpt-5`: 81.2%) while others struggle (`qwen3-vl-32b-instruct`: 18.8%). More nuanced dimensions—uniqueness, stability, existence—also prove challenging, with uniqueness showing extreme variance (1.2–95.0%) across models.

The equilibrium scene, time characteristic, and time horizon dimensions show the greatest performance spread. These require understanding subtle distinctions in how economic concepts are defined and applied across different historical contexts. The low scores suggest that temporal reasoning and historical concept evolution are particularly difficult for current LLMs.

Interestingly, GPT-5 series models show unusually weak CI performance relative to their strong performance on other tasks. This may indicate that the training data or optimization objectives for these models prioritize contemporary language patterns over historical economic terminology, creating a systematic weakness in this dimension.

G Task Generation Methodology

This appendix provides comprehensive documentation of the multi-round generation pipeline used to construct evaluation items for each task dimension in AER-bench. All tasks except Concept Identification (CI) employ LLM-based generation with carefully designed prompts and validation steps. CI items are manually annotated from historical AER papers and are therefore not covered in this appendix.

For each task, we describe: (1) the complete multi-round generation workflow, (2) key design rationale, (3) full prompt templates with all critical instructions, (4) complete output format specifications, and (5) scoring rules and evaluation criteria. All generation uses `gpt-4o` as the extraction model.

G.1 Economic Knowledge (EK)

G.1.1 Task Overview

Economic Knowledge (EK) is a Level 1 (foundational capability) task that evaluates models' depth of understanding of economic literature. The core evaluation goal is to test whether a model can answer questions about specific papers' model setups, core findings, and methodological choices. Each paper produces 20 objective questions: 5 single-choice, 5 true/false, 5 fill-in-blank, and 5 ordering questions. The task depends only on PDF input (no supplementary materials required) and uses a 2-round generation pipeline.

G.1.2 Generation Workflow

EK employs a two-round pipeline to ensure balanced knowledge coverage and question diversity:

1. **Round 1 (PDF input)**: Extract 12–15 paper-specific knowledge points spanning six categories: model setup, findings, methods, mechanisms, contributions, and institutional context. Each knowledge point is tagged as *quantitative* (involving specific numbers) or *conceptual* (involving qualitative understanding), with a target of at least 50% conceptual points to avoid over-extraction of numerical facts.
2. **Round 2 (text input)**: Based on the structured knowledge point list from Round 1, generate 20 objective questions: 5 single-choice, 5 true/false, 5 fill-in-blank, and 5 ordering questions. Each question references the source knowledge point and includes difficulty tags.

G.1.3 Design Rationale

Separating knowledge extraction from question generation prevents two common failure modes: (1) knowledge point omission when simultaneously reading the paper and generating 20 questions, and (2) uneven distribution across question types or content categories. The structured knowledge point list serves as a checklist to ensure comprehensive coverage.

Critical Design Constraint All questions are designed to test whether an AI system or economics researcher **knows this specific paper**. Questions provide ONLY the first author's surname + year + a short topic phrase (e.g., "Autor et al. (2013), a paper on Chinese import competition and U.S. labor markets"). The full paper title, journal name, and co-author details do NOT appear in the stem. The answer must require knowledge of the paper's specific details that cannot be guessed from the topic phrase alone.

G.1.4 Round 1 Prompt Template

```
You are a professor of economics. Please read the attached PDF paper
in full and extract **paper-specific knowledge points** that can be
used for exam questions.

### Critical design constraint

The exam questions you will later create are designed to test whether
an AI system or an economics researcher **knows this specific paper**.
Therefore:
```

```
- **ALL knowledge points must be about this paper's specific content**:
  its model setup, assumptions, variables, findings, methodology,
  mechanisms, contributions, or institutional context.
- The questions will provide ONLY the **first author's surname + year
  + a short topic phrase** (e.g., "Autor et al. (2013), a paper on
  Chinese import competition and U.S. labor markets"). The full paper
  title, journal name, and co-author details will NOT appear in the
  stem. The **answer must require knowledge of the paper's specific
  details** that cannot be guessed from the topic phrase alone.
- A well-read economist who has studied this paper (or can search for
  it) should be able to answer; one who has not should find it difficult.

### Knowledge point categories

Extract knowledge points spanning as many of the following categories
as the paper supports:

- `model_setup`: Specific assumptions, agent definitions, variable
  choices, functional forms, or information structures in the paper's
  model
- `finding`: Core empirical or theoretical results -- coefficient
  directions, magnitudes, significance levels, comparative statics,
  or proposition statements
- `method`: Identification strategy, estimation approach, instrument
  choice, data sources, or sample construction decisions and their
  rationale
- `mechanism`: Economic mechanisms or causal channels that the paper
  identifies, tests, or discusses
- `contribution`: The paper's stated contributions, how it differs
  from or extends prior work
- `institution`: Institutional background, policy details, or market
  features specific to the paper's setting

### CRITICAL: Content type balance

You MUST extract a balanced mix of **quantitative** and **conceptual**
knowledge points:

- **Quantitative** knowledge points: involve specific numbers,
  coefficients, percentages, counts, or magnitudes from the paper
  (e.g., "The estimated effect is -0.37", "The sample includes 126 DMAs")
- **Conceptual** knowledge points: involve model assumptions, mechanism
  descriptions, methodology choices, variable definitions, institutional
  features, or qualitative findings (e.g., "The paper uses a shift-share
  instrument", "The model assumes risk-neutral agents")

**Target: at least 50% of knowledge points must be conceptual.** Do not
over-extract numerical facts at the expense of conceptual understanding.
```

G.1.5 Round 1 Output Format

```
{
  "paper_id": "<paper ID>",
  "paper_title": "<paper title>",
  "knowledge_points": [
    {
      "kp_id": "KP01",
      "category": "model_setup | finding | method | mechanism |
                  contribution | institution",
      "content_type": "quantitative | conceptual",
      "statement": "<specific knowledge point>",
      "paper_context": "<2-3 sentence description stating what aspect
                        of the paper this addresses, the specific detail
                        tested, and where it appears>"
```

```
    }
  ]
}
```

G.1.6 Round 2 Prompt Template

```
You are a professor of economics. Below is a list of paper-specific
knowledge points extracted from an economics paper. Based on these
knowledge points, generate **20 objective questions** to test whether
an examinee knows this specific paper.

### Question type requirements

Generate exactly:
- 5 single-choice questions (4 options each, labeled A/B/C/D)
- 5 true/false questions
- 5 fill-in-the-blank questions
- 5 ordering questions (4 items to order)

### CRITICAL: Content type balance

Ensure that across all 20 questions, at least 50% test **conceptual**
knowledge (not just numerical recall). Use the content_type tags from
the knowledge points to guide this balance.

### Mandatory stem format

Every question stem MUST begin with the paper reference in this exact
format:

"In [FirstAuthor] et al. ([Year]), a paper on [brief topic phrase], ..."

Example: "In Autor et al. (2013), a paper on Chinese import competition
and U.S. labor markets, what identification strategy is used?"

Do NOT include:
- Full paper title
- Journal name
- Full author list beyond first author

### Question design principles

1. **Paper-specific**: Every question must require knowledge of this
   specific paper's content. Avoid generic economics questions.
2. **Unambiguous**: Each question must have a single, clearly correct
   answer based on the paper.
3. **Difficulty distribution**: Aim for a mix of medium (60%) and hard
   (40%) questions. Easy questions that can be guessed should be avoided.
4. **Source tracking**: Each question must reference its source
   knowledge point via source_kp field.

### Single-choice question requirements

- Provide exactly 4 options (A, B, C, D)
- All distractors must be plausible but clearly incorrect based on the
  paper
- Avoid "all of the above" or "none of the above" options
- Distractors should not be trivially eliminable

### True/false question requirements

- Statement must be specific enough that truth value is unambiguous
- Avoid statements that are "technically true but misleading" or vice
  versa
- Explanation must cite specific paper content
```

```
### Fill-in-the-blank question requirements

- Blank should test a specific term, number, or concept
- Use "___" to mark the blank position
- Provide canonical answer and list of acceptable synonyms/variants
- Acceptable answers should account for: capitalization differences,
  singular/plural forms, common abbreviations

### Ordering question requirements

- Provide exactly 4 items to be ordered
- Specify the ordering criterion clearly in the stem (e.g., "from
  smallest to largest effect size", "in chronological order of
  implementation")
- Items must have an unambiguous correct ordering based on the paper
- Answer field contains the items in correct order
```

G.1.7 Round 2 Output Format

```
{
  "paper_id": "<paper ID>",
  "questions": [
    {
      "id": "EK_<paper_id>_SC01",
      "type": "single_choice",
      "content_type": "conceptual | quantitative",
      "source_kp": "KP03",
      "difficulty": "medium | hard",
      "knowledge_tags": ["<knowledge category>"],
      "stem": "<question stem starting with paper reference>",
      "options": {
        "A": "<option A>",
        "B": "<option B>",
        "C": "<option C>",
        "D": "<option D>"
      },
      "answer": "B",
      "explanation": "<explanation referencing specific paper content>"
    },
    {
      "id": "EK_<paper_id>_TF01",
      "type": "true_false",
      "content_type": "conceptual | quantitative",
      "source_kp": "KP07",
      "difficulty": "medium | hard",
      "knowledge_tags": ["<knowledge category>"],
      "stem": "<statement about the paper to judge true/false>",
      "answer": true,
      "explanation": "<explanation referencing specific paper content>"
    },
    {
      "id": "EK_<paper_id>_FB01",
      "type": "fill_in_blank",
      "content_type": "conceptual | quantitative",
      "source_kp": "KP05",
      "difficulty": "medium | hard",
      "knowledge_tags": ["<knowledge category>"],
      "stem": "<fill-in stem with ___ referencing paper content>",
      "answer": "<canonical answer>",
      "acceptable_answers": ["<synonym 1>", "<synonym 2>"],
      "explanation": "<explanation referencing specific paper content>"
    },
    {
      "id": "EK_<paper_id>_OD01",
```

```
      "type": "ordering",
      "content_type": "conceptual | quantitative",
      "source_kp": "KP08",
      "difficulty": "medium | hard",
      "knowledge_tags": ["<knowledge category>"],
      "stem": "<ordering stem stating the criterion, referencing paper>",
      "items": ["<item A>", "<item B>", "<item C>", "<item D>"],
      "answer": ["<item B>", "<item A>", "<item D>", "<item C>"],
      "explanation": "<explanation referencing specific paper content>"
    }
  ]
}
```

G.1.8 Scoring Rules

Question Type	Fully Correct	Partially Correct	Incorrect
Single Choice (5 questions)	1 point	–	0 points
True/False (5 questions)	1 point	–	0 points
Fill-in-Blank (5 questions)	1 point (case-insensitive, or in acceptable_answers)	–	0 points
Ordering (5 questions)	1 point (exact match)	0.5 points (Kendall $\tau \geq 0.5$)	0 points

Total score = sum of individual question scores / number of questions.

G.2 Mathematical Modeling (MM)

G.2.1 Task Overview

Mathematical Modeling (MM) is a Level 1 (foundational capability) task that evaluates models' ability to write reasonable model equations, constraints, or complete key derivation steps given an economic scenario. The core evaluation goal is to test formula-level mathematical skills. Each paper produces 3 questions spanning three types: equation design, formula derivation, and equilibrium solving. The task depends only on PDF input and uses a 2-round generation pipeline.

G.2.2 Generation Workflow

MM uses a two-round pipeline to systematically extract and select mathematical content:

1. **Round 1 (PDF input)**: Systematically extract all mathematical modeling content from the paper, including model setup (agents, variables, parameters), key equations (objective functions, constraints, FOCs), propositions and proofs, and derivation chains. For empirical papers, extract regression equations and identification assumptions. Output includes a `suitable_for_mm` flag.
2. **Round 2 (text input)**: Based on Round 1's structured mathematical content, design 3 self-contained mathematical modeling questions spanning three types: equation design, formula derivation, and equilibrium solving. Each question includes scenario description, notation conventions, known conditions, and standard answer in LaTeX.

G.2.3 Design Rationale

Economic papers scatter mathematical content across main text, appendices, and footnotes. Round 1 performs exhaustive extraction and builds derivation chains, allowing Round 2 to select the most suitable elements for exam questions without missing key formulas or choosing overly trivial derivations.

LaTeX Escaping in JSON Since output is JSON containing LaTeX strings, all backslashes must be escaped: `\frac{a}{b}` becomes `\\frac{a}{b}` in JSON. Every backslash \ in LaTeX must be written as \\ in the JSON string. Double backslashes in LaTeX (e.g., \\ for line breaks) must be \\\\ in JSON.

G.2.4 Round 1 Prompt Template

```
You are an expert in mathematical economics. Read the full attached
PDF paper and systematically extract the paper's mathematical modeling
content.

### Extraction requirements

Identify and record the following:

1. **Model setup**: Definitions of all economic agents, variables, and
   parameters
2. **Key equations**: Objective functions, constraints, equilibrium
   conditions, first-order conditions, etc.
3. **Propositions and proofs**: Propositions, Theorems, and Lemmas in
   the paper and their core derivation steps
4. **Derivation chains**: Logical relations between formulas (which
   formula follows from which prior conditions)

### Handling non-theoretical papers

If the paper is empirical:
- Extract regression / estimation equations (e.g. DID, IV, structural
  models, etc.)
- Extract mathematical formulations of identification assumptions
- Extract statistical or asymptotic properties of the model

If the paper contains almost no mathematical derivations suitable for
exam items (e.g. only reports regression results, lacking
reconstructible equations and derivation chains):
- Set `suitable_for_mm = false`
- Provide `reason_not_suitable`
- Do not invent mathematical models or derivation chains

### LaTeX in JSON -- escaping rules

Since the output is JSON containing LaTeX strings, you MUST follow
these escaping rules:
- Every backslash `\` in LaTeX must be written as `\\` in the JSON
  string
- For example: `\frac{a}{b}` must be written as `\\frac{a}{b}`
- Newlines in LaTeX should be `\\n` in the JSON string
- Double backslashes in LaTeX (e.g., `\\` for line breaks) must be
  `\\\\` in JSON
- Test: the JSON must be parseable by a standard JSON parser
```

G.2.5 Round 1 Output Format

```
{
  "paper_id": "<paper ID>",
  "paper_title": "<paper title>",
  "suitable_for_mm": true,
  "reason_not_suitable": "",
  "paper_type": "theoretical | empirical | mixed",
  "model_elements": {
    "agents": [
      {"name": "<agent>", "variables": ["<variable (LaTeX)>"],
       "parameters": ["<parameter>"]}
    ],
    "key_equations": [
      {
        "eq_id": "EQ01",
        "label": "<equation label, e.g. objective function, budget
                  constraint, FOC>",
        "latex": "<LaTeX formula>",
```

```
        "context": "<where the equation appears and its meaning
                    (1 sentence)>",
        "depends_on": ["EQ00"]
      }
    ],
    "propositions": [
      {
        "prop_id": "P01",
        "statement": "<verbal statement of the proposition>",
        "key_result_latex": "<LaTeX expression of the core result>",
        "proof_sketch": "<key proof steps (2-3 steps)>",
        "depends_on": ["EQ01", "EQ03"]
      }
    ]
  },
  "candidate_questions": [
    {
      "suggested_type": "equation_design | derivation |
                         equilibrium_solving",
      "target_element": "EQ03 | P01",
      "why_suitable": "<why this element is suitable as an exam item
                       (1 sentence)>",
      "estimated_difficulty": "medium | hard"
    }
  ]
}
```

G.2.6 Round 2 Prompt Template

```
You are an expert in writing mathematical economics exam questions.
Below is extracted mathematical modeling content from an economics
paper. Based on this content, design **3 mathematical modeling
questions**.

### Question type requirements

Design exactly 3 questions, one from each type:
1. **Equation design**: Given a scenario, write the appropriate
   objective function, constraint, or equilibrium condition
2. **Formula derivation**: Derive a key result from given premises
3. **Equilibrium solving**: Characterize equilibrium or solve for
   optimal choices

### Self-containment principle

Each question must be **completely self-contained**:
- Provide all necessary background and notation in the question stem
- Define all variables and parameters used
- State all assumptions and known conditions
- The examinee should NOT need to refer to the original paper

### Question structure

Each question should include:
1. **Scenario description**: Economic context and setup (2-3 sentences)
2. **Notation conventions**: List of all symbols with their meanings
3. **Known conditions**: What is given or assumed
4. **Task**: What the examinee needs to derive or solve
5. **Standard answer**: Complete solution in LaTeX

### LaTeX escaping reminder

Remember: every `\` in LaTeX must be `\\` in JSON output.

### Difficulty calibration
```

```
- **Medium**: Requires applying standard techniques (FOC, envelope
  theorem, budget constraint substitution)
- **Hard**: Requires multi-step reasoning, non-obvious substitutions,
  or careful handling of corner cases

Target: 1-2 medium questions, 1-2 hard questions.
```

G.2.7 Round 2 Output Format

```
{
  "paper_id": "<paper ID>",
  "questions": [
    {
      "id": "MM_<paper_id>_Q01",
      "type": "equation_design | derivation | equilibrium_solving",
      "difficulty": "medium | hard",
      "scenario": "<economic scenario description (2-3 sentences)>",
      "notation": {
        "<symbol>": "<meaning>",
        "x": "consumption quantity",
        "p": "price"
      },
      "known_conditions": [
        "<condition 1>",
        "<condition 2>"
      ],
      "task": "<what the examinee needs to do>",
      "standard_answer": {
        "latex": "<complete LaTeX solution>",
        "explanation": "<step-by-step explanation>"
      },
      "rubric": {
        "correctness": "<what constitutes a correct answer>",
        "key_elements": ["<element 1>", "<element 2>"]
      }
    }
  ]
}
```

G.2.8 Scoring Method

MM uses a hybrid scoring approach combining structured comparison and LLM-as-Judge:

Scoring Dimensions (0–10 points each, total 40)

1. **Formula correctness** (0–10): All formulas are mathematically equivalent to the reference; derivation steps are valid. 9–10: fully correct; 6–8: minor errors but core structure correct; 3–5: significant errors but partial understanding evident; 0–2: fundamentally incorrect.
2. **Notation consistency** (0–10): All symbols match the problem setup and are used consistently. 9–10: perfect consistency; 6–8: minor inconsistencies; 3–5: some symbols undefined or inconsistent; 0–2: major notation confusion.
3. **Derivation completeness** (0–10): All intermediate steps shown; a reader can verify each step. 9–10: all steps present; 6–8: key steps present but some algebra omitted; 3–5: major steps missing; 0–2: only final answer or fragmentary.
4. **Economic intuition** (0–10): Clear interpretation of results in economic terms; explains what the math means. 9–10: excellent economic interpretation; 6–8: some interpretation provided; 3–5: minimal economic context; 0–2: no economic interpretation.

Normalized score = total / 40.

G.3 Paper Abstraction (PA)

G.3.1 Task Overview

Paper Abstraction (PA) is a Level 2 (comprehensive capability) task that evaluates models' ability to write abstracts after reading condensed body text. The core evaluation goal is to test whether generated abstracts cover key points from the original abstract. Each paper produces 1 evaluation item containing condensed body text (approximately 30% of original length, excluding abstract) and 5–10 atomic semantic propositions as evaluation criteria. The task depends only on PDF input and uses a 3-round generation pipeline.

G.3.2 Generation Workflow

PA employs a three-round pipeline to produce condensed text and atomic evaluation criteria:

1. **Round 1 (PDF input)**: Extract the original abstract verbatim, identify section structure, and summarize core arguments for each major section (1–2 sentences per section). Also extract 2–4 claimed contributions.
2. **Round 2 (PDF + Round 1)**: Generate condensed body text at approximately 30% of the original length, excluding the abstract. Use Round 1's section structure to guide selective retention of core arguments, key findings, and methodological details while removing redundant exposition.
3. **Round 3 (text input)**: Decompose the original abstract into 5–10 atomic semantic propositions (key points) that serve as evaluation criteria. Cross-validate these propositions against Round 1's section arguments to ensure completeness.

G.3.3 Design Rationale

The three-round separation ensures: (1) Round 1 provides structural guidance for Round 2's condensation, (2) Round 2 can refer back to the PDF to avoid incorrect deletions of critical arguments, and (3) Round 3 focuses purely on atomic proposition decomposition without needing PDF access.

Examinee Input Examinees receive only the condensed body text (approximately 30% of original length, excluding abstract) in plain text format. This design reduces token cost, supports models without PDF input, and focuses evaluation on "distilling abstracts from core information" rather than "processing long documents."

G.3.4 Round 1 Prompt Template

```
You are an expert in economics research methods. Please read the
attached PDF paper and complete the following information extraction.

### Task 1: Extract the original abstract

Locate and extract the paper's abstract text in full (verbatim copy;
do not paraphrase).

### Task 2: Paper structure and core arguments by section

Identify the paper's section structure and summarize the core argument
in 1-2 sentences for each major section.

### Task 3: Identify the paper's core contributions

List 2-4 main contributions the paper explicitly claims (typically at
the end of the introduction or in the conclusion).
```

G.3.5 Round 1 Output Format

```
{
  "paper_id": "<paper ID>",
```

```
  "paper_title": "<paper title>",
  "original_abstract": "<full original abstract, verbatim>",
  "sections": [
    {
      "section_number": "1",
      "section_title": "<section title>",
      "core_argument": "<core argument (1-2 sentences)>"
    }
  ],
  "claimed_contributions": ["<contribution 1>", "<contribution 2>"],
  "methodology_summary": "<one-sentence overview of methods>",
  "key_findings_summary": "<one-sentence overview of findings>"
}
```

G.3.6 Round 2 Prompt Template

```
You are an academic editor. Below is the section structure extracted
from an economics paper. Please re-read the attached PDF source and
compress the body of the paper into a **condensed version of
approximately 30% of the original length**.

### Compression goal

Produce a **plain-text condensed body** whose length is approximately
30% of the original paper body (excluding references and appendices).
This text will be the sole input for examinees, who will write an
abstract based on it.

### Compression principles

1. **Must not include the original abstract** -- that is what examinees
   need to produce
2. **Preserve the paper's core argumentative chain**: research question
   -> methods -> key findings -> conclusions
3. **Organize in section order**, retaining the 2-4 most critical
   sentences of core argument per section
4. **Retain key data and quantitative results** (e.g., core regression
   coefficients, effect sizes, statistical significance)
5. **Retain core methodological description** (e.g., identification
   strategy, model specification, data sources)
6. **Omit**: detailed literature review, secondary variants of
   robustness checks, appendix material, redundant exposition
7. **Do not paraphrase** -- prefer direct quotations or minimally
   shortened phrases from the original
8. Use **English** and maintain an academic register
9. Prefix each section paragraph with `[Section N: Title]` for clarity
```

G.3.7 Round 3 Prompt Template

```
You are an expert in academic writing. Below is the original abstract
of an economics paper. Decompose it into **5-10 atomic semantic
propositions** (key points) that will serve as evaluation criteria for
examinee-generated abstracts.

### Decomposition principles

1. **Atomic**: Each key point should express ONE distinct claim or fact
2. **Verifiable**: Each point should be checkable against candidate text
3. **Complete coverage**: Together, the points should cover all major
   elements of the original abstract
4. **Category balance**: Include points from multiple categories:
   research question, method, finding, contribution

### Key point categories
```

```
- ‘research_question‘: What economic question the paper addresses
- ‘method‘: Identification strategy, data, or modeling approach
- ‘finding‘: Core empirical or theoretical results
- ‘contribution‘: How the paper advances the literature

### must_appear flag

Mark a key point as ‘must_appear: true‘ if it represents a core
contribution or finding that MUST be mentioned in any reasonable
abstract of this paper. Mark as ‘false‘ if it is a supporting detail
that could reasonably be omitted in a concise abstract.

Target: 60-70% of key points should be must_appear: true.
```

G.3.8 Round 3 Output Format

```
{
  "paper_id": "<paper ID>",
  "key_points": [
    {
      "kp_id": "KP01",
      "statement": "<atomic proposition>",
      "category": "research_question | method | finding | contribution",
      "must_appear": true
    }
  ]
}
```

G.3.9 Scoring Method

PA uses LLM-as-judge evaluation across four dimensions (0–10 points each):

1. **Coverage** (0–10): Are all or nearly all reference key points mentioned or clearly implied? 9–10: all key points covered; 6–8: most covered with 1–2 minor omissions; 3–5: several important points missing; 0–2: only 1–2 points covered.
2. **Accuracy** (0–10): Are all stated claims factually correct per the paper content? 9–10: all correct; 6–8: minor inaccuracies; 3–5: some factual errors; 0–2: major errors.
3. **Conciseness** (0–10): Is the abstract within word limit with no redundancy? 9–10: within limit, every sentence adds value; 6–8: mostly concise with minor redundancy; 3–5: noticeably wordy or exceeds limit; 0–2: excessively long.
4. **Academic style** (0–10): Does it read like a published abstract in a top economics journal? 9–10: publication-ready; 6–8: generally academic with some informal phrasing; 3–5: mixed register; 0–2: not academic.

G.4 Literature Review (LR)

G.4.1 Task Overview

Literature Review (LR) is a Level 2 (comprehensive capability) task that evaluates models' ability to write introductions with citation networks consistent with the original paper. The core evaluation goal is to test whether models can produce literature reviews that align with the paper's actual citation structure. Each paper produces 1 evaluation item containing condensed body text (approximately 50% of original length, excluding introduction and references), three-tier citation classification, citation role annotations, and a paper-specific writing rubric. The task depends only on PDF input and uses a 4-round generation pipeline.

G.4.2 Generation Workflow

LR uses a four-round pipeline to construct citation-aware evaluation items:

1. **Round 1 (PDF input)**: Extract the complete reference list from the paper, including author names, titles, publication venues, and years. Tag each reference by publication type (journal article, book, working paper, etc.).
2. **Round 2 (PDF + Round 1)**: Analyze the introduction section: identify introduction boundaries, extract the argumentation structure, and annotate each citation's role in the argument. Classify citations into three tiers based on journal quality and relevance to the paper's core argument.
3. **Round 3 (PDF + Round 2)**: Generate condensed body text at approximately 50% of original length, excluding the introduction and references. Use Round 2's introduction boundaries to ensure precise exclusion.
4. **Round 4 (text input)**: Construct a paper-specific writing rubric based on the argumentation structure identified in Round 2. The rubric captures the paper's specific rhetorical strategy rather than applying a generic template.

G.4.3 Design Rationale

The four-round separation ensures: (1) clean reference list extraction before citation role annotation, (2) precise introduction exclusion using Round 2 boundaries, and (3) customized rubric construction based on the paper's specific argumentation strategy.

Citation Tier System Citations in the introduction are classified into three tiers based on journal quality and relevance:

- **Tier 1 (Must-cite)**: Core references indispensable to the argument, typically from top journals (AER, Econometrica, QJE, JPE, REStud) or high-impact sources, but not limited to these—any high-impact work critical to the argument qualifies.
- **Tier 2 (Should-cite)**: Relevant formal academic publications from reputable but non-top journals, academic books, or well-known working papers (NBER, CEPR).
- **Tier 3 (Optional)**: Lower-relevance or less authoritative sources (obscure journals, general preprints, technical reports).

Scoring Formula Weighted citation recall: $0.8 \times$ Tier1_recall $+ 0.2 \times$ Tier2_recall (Tier 3 not scored). Final score: 50% citation recall + 50% writing rubric.

G.4.4 Round 1 Prompt Template

```
You are an expert in academic reference management. Read the attached
PDF paper and extract the **complete reference list** from the
References / Bibliography section.

### Extraction requirements

For each reference, extract:
- Authors (standard format)
- Year
- Title
- Journal / publisher
- Publication type (venue_type)
- A standardized citation_key

### venue_type categories

Label each item by its publication source:
- ‘top5‘: economics Top 5 journals (AER, Econometrica, QJE, JPE, REStud)
- ‘top_field‘: widely recognized top field journals (e.g., JF, JFE,
  RFS, JME, JEEA, AEJ series, JHE, JDE, JUE, JPubE, JLE, RAND, EJ,
  REE, etc.)
- ‘good_journal‘: well-known but non-top-tier formal journals
- ‘book‘: scholarly monographs or book chapters
- ‘conference‘: conference papers
```

```
- `working_paper_top`: working papers from prominent institutions
  (NBER, CEPR, World Bank, IMF, Fed, etc.)
- `working_paper_other`: working papers from other institutions
- `preprint`: preprints (arXiv, SSRN, etc.)
- `other`: reports, technical documents, news, web pages, etc.
```

G.4.5 Round 2 Prompt Template

```
You are an expert in economics research writing. Re-read the attached
PDF paper and analyze the **introduction section** in detail.

### Task 1: Identify introduction boundaries

Precisely identify where the introduction begins and ends (page numbers
and section numbers).

### Task 2: Extract argumentation structure

Identify the major argumentative modules in the introduction (e.g.,
motivation, literature gap, methodology preview, contribution summary).
For each module, provide a 2-3 sentence summary.

### Task 3: Annotate each citation's role

For every citation that appears in the introduction, record:
- citation_key (matching Round 1 format)
- The argumentative role it plays (e.g., "establishes empirical
  regularity", "represents prior theoretical approach", "identifies
  literature gap")
- Tier classification (1/2/3) based on journal quality and relevance

### Tier classification criteria

**Tier 1 (Must-cite)**: Core references indispensable to the argument.
Typically from top journals (AER, Econometrica, QJE, JPE, REStud) or
high-impact sources, but not limited to these---any high-impact work
critical to the argument qualifies.

**Tier 2 (Should-cite)**: Relevant formal academic publications from
reputable but non-top journals, academic books, or well-known working
papers (NBER, CEPR).

**Tier 3 (Optional)**: Lower-relevance or less authoritative sources.

### IMPORTANT: Tier 1 and Tier 2 count requirements

After classification, verify:
- Tier 1 should have >= 10 citations
- Tier 2 should have >= 15 citations

If counts are insufficient, re-examine the introduction and promote
relevant citations from lower tiers.
```

G.4.6 Round 3 Prompt Template

```
You are an academic editor. Below is the introduction boundary
information extracted from an economics paper. Re-read the attached
PDF and compress the body of the paper (EXCLUDING the introduction
and references) into approximately 50% of the original length.

### Compression goal

Produce plain-text condensed body at ~50% of original length, excluding
introduction and references. This will be the examinee's input for
writing an introduction.
```

```
### Compression principles

1. **Must exclude the entire introduction section** (use Round 2
   boundaries)
2. **Must exclude the references section**
3. **Preserve core argumentative chain**: methods -> results ->
   robustness -> conclusion
4. **Retain key quantitative results** and methodological details
5. **Omit**: detailed proofs, secondary robustness checks, appendix
   material
6. Use section prefixes `[Section N: Title]` for clarity
```

G.4.7 Round 4 Prompt Template

```
You are an expert in economics research writing. Based on the
introduction analysis from Round 2, construct a **paper-specific
writing rubric** that captures this paper's actual rhetorical strategy.

### Rubric construction principles

1. **Paper-specific dimensions**: Identify 4-6 dimensions based on what
   THIS paper's introduction actually does (not a generic template)
2. **Concrete evaluation criteria**: For each dimension, list specific
   elements to check
3. **Weight allocation**: Assign weights summing to 100 based on
   importance
4. **Tier-classified citation list**: Aggregate citation_keys by tier
   from Round 2

### Example dimensions (adapt to paper)

- Motivation and research question clarity
- Literature gap identification
- Methodology preview
- Contribution summary
- Institutional context (if applicable)
- Theoretical framework setup (if applicable)

### Output requirements

- writing_rubric: array of dimension objects with name, description,
  key_elements_to_check, weight
- gold_citations_tiered: object with tier_1, tier_2, tier_3 arrays
- scoring_formula: explicit formula for combining citation and rubric
  scores
```

G.4.8 Scoring Method

LR uses dual-track scoring:

Citation Recall (50% of final score)

```
tier1_recall = |tier1_gold  pred_set| / |tier1_gold|
tier2_recall = |tier2_gold  pred_set| / |tier2_gold|
weighted_recall = 0.8 × tier1_recall + 0.2 × tier2_recall
citation_score = weighted_recall × 100
```

Design rationale: Use Recall (not F1) because introduction writing is open-ended; examinees may reasonably cite works outside the gold set. Tier 1 (80% weight) represents indispensable core references; Tier 2 (20% weight) represents quality supplementary references. Tier 3 not scored.

Writing Rubric (50% of final score) LLM-as-judge evaluates the introduction against paper-specific rubric dimensions. Each dimension scored 0–10, weighted by dimension weights, normalized to 0–100 scale.

Final Score

```
final_score = 0.5 × citation_score + 0.5 × rubric_score
```

G.5 Method Generation (MG)

G.5.1 Task Overview

Method Generation (MG) is a Level 2 (comprehensive capability) task that evaluates models' ability to construct economic models from natural language scenarios. The core evaluation goal is to test whether models can clearly specify economic agents, optimization problems, constraints, and equilibrium/game mechanisms. Each paper produces 1 evaluation item containing a desymbolized scenario narrative, standard model structure, and element-counting rubric. The task depends only on PDF input and uses a 3-round generation pipeline.

G.5.2 Generation Workflow

MG employs a three-round pipeline to transform formal models into natural language scenarios:

1. **Round 1 (PDF input)**: Fully extract the paper's core economic model, including all agents and their choice variables, objective functions (with optimization direction), all constraints (budget, resource, IC, PC, feasibility), equilibrium concept, key FOCs, and core mechanisms. For empirical papers, extract the theoretical framework or structural model if present.
2. **Round 2 (PDF + Round 1)**: Rewrite Round 1's formal model into a desymbolized natural language scenario description. Remove all mathematical notation and variable symbols, replacing them with plain economic intuition and verbal descriptions of relationships.
3. **Round 3 (text input)**: Construct an element-counting rubric based on Round 1's model structure. Dynamically allocate points across dimensions (agents, objectives, constraints, equilibrium characterization) based on the actual number of elements in the model.

G.5.3 Design Rationale

Round 1 focuses purely on faithful extraction without rewriting. Round 2 can then generate targeted natural language descriptions while referring back to the PDF to ensure economic intuition accuracy. Round 3 is purely computational—allocating rubric points based on element counts—and requires no PDF access.

G.5.4 Round 1 Prompt Template

```
You are an economic theory expert. Read the attached PDF paper and
**fully extract** the paper's core economic model.

### Extraction requirements

Strictly extract the following from the paper as written (do not omit
any agent or constraint):

1. **Economic agents**: All agents in the model and their choice
   variables
2. **Objective functions**: Each agent's optimization objective
   (including maximization/minimization direction)
3. **Constraints**: All constraints each agent faces (budget, resource,
   incentive compatibility, participation, feasibility, etc.)
4. **Equilibrium concept**: The equilibrium / solution concept used
5. **Key first-order conditions**: FOCs / characterization results as
   derived or stated
6. **Core mechanism**: Economic forces that drive the model's results
```

```
### Handling non-theoretical papers

If the paper is empirical:
- Extract the theoretical framework or structural model it relies on
  (if any)
- Extract the economic logic behind the identification strategy
- If the paper has no theoretical model at all, state that the paper
  is unsuitable for the MG task

When `suitable_for_mg = false`:
- Provide `reason_not_suitable`, briefly explaining why
- Other model fields may be empty arrays / empty objects (do not
  fabricate a model)
```

G.5.5 Round 2 Prompt Template

```
You are an economics textbook author. Below is the full model structure
extracted from an economics paper. Rewrite it as a **desymbolized
economic scenario narrative** for examinees to model from.

### IMPORTANT: Check suitability first

If the Round 1 output has `suitable_for_mg = false`, output a minimal
JSON indicating the paper is unsuitable and stop.

### Desymbolization requirements

1. **Remove ALL mathematical symbols and variable names**: No x, y, p,
   w, , , etc.
2. **Use plain economic language**: "consumption quantity" not "x",
   "wage rate" not "w"
3. **Describe relationships verbally**: "utility increases with
   consumption" not "U(x) is increasing"
4. **Preserve economic intuition**: The scenario must convey the same
   economic trade-offs as the formal model
5. **Self-contained**: Examinee should NOT need the original paper

### Scenario structure

1. **Economic context** (2-3 sentences): What is the economic
   environment?
2. **Agents and their goals** (1 sentence per agent): Who are the
   decision-makers and what do they want?
3. **Constraints and trade-offs** (2-3 sentences): What limits their
   choices?
4. **Equilibrium/outcome concept** (1 sentence): What constitutes a
   solution?

### Examinee task description

After the scenario, provide clear instructions: "Based on the scenario
above, construct a formal economic model that captures the key
trade-offs and decision problems. Your model should specify: (1) all
economic agents and their choice variables, (2) each agent's objective
function, (3) all relevant constraints, (4) the equilibrium concept."
```

G.5.6 Round 3 Prompt Template

```
You are an exam rubric designer. Based on the model structure extracted
in Round 1, construct an **element-counting rubric** that dynamically
allocates points based on the actual number of model elements.

### Rubric construction algorithm
```

```
Total points: 100

1. **Agent identification** (15-25 points): Allocate based on number
   of agents. If 1 agent: 15 pts. If 2 agents: 20 pts. If 3+ agents:
   25 pts.

2. **Objective specification** (20-30 points): Allocate based on
   complexity. Simple single-period utility: 20 pts. Multi-period or
   strategic objectives: 25-30 pts.

3. **Constraint completeness** (30-40 points): Allocate proportionally
   to number of constraints. Each constraint type worth 5-10 pts
   depending on complexity.

4. **Equilibrium characterization** (15-20 points): Standard
   equilibrium concept: 15 pts. Complex equilibrium (e.g., perfect
   Bayesian): 20 pts.

### Quality levels for each element

- **correct** (full weight): Element present, correctly formalized,
  economically meaningful
- **substantial** (70% weight): Mostly correct with minor issues
- **partial** (40% weight): Attempted but significant issues
- **minimal** (15% weight): Vaguely referenced without proper
  formalization
- **missing** (0 points): Not present
```

G.5.7 Scoring Method

MG uses element-counting rubric (total 100 points) with LLM-assisted validation of element correctness. Points are dynamically allocated based on model complexity. For each rubric element, LLM judges quality level (correct/substantial/partial/minimal/missing) and awards corresponding percentage of per-element points. Total score is sum of all element scores.

G.6 Economic Analysis (EA)

G.6.1 Task Overview

Economic Analysis (EA) is a Level 2 (comprehensive capability) task that evaluates models' ability to write professional, multi-dimensional economic analysis given data results or model outputs. The core evaluation goal is to test whether models can produce substantive economic interpretations of empirical or theoretical results. Each paper produces 1–2 evaluation items (one per core result), each containing a purely factual data description, a reference analysis paragraph, applicable scoring dimensions, and examinee instructions. The task depends only on PDF input and uses a 2-round generation pipeline.

G.6.2 Generation Workflow

EA uses a two-round pipeline to separate data description from analytical writing:

1. **Round 1 (PDF input)**: Scan all tables and figures in the paper, label their type and importance, and select 1–2 core results that best represent the paper's contribution (typically main regression tables or core theoretical result figures). For each selected result, extract a context paragraph (150–250 words covering research question, data, methodology, and what the result represents) and a purely factual data description (200–400 words) without any interpretation or analysis.
2. **Round 2 (PDF + Round 1)**: For each result identified in Round 1, locate the corresponding analysis paragraph(s) in the paper's main text (400–600 words). Extract the author's original analysis as the gold standard. Identify which of the six evaluation dimensions apply to this specific result.

G.6.3 Design Rationale

The core challenge is cleanly separating "result description" from "analysis paragraph." Round 1 identifies core results and extracts pure data descriptions (no interpretation), then Round 2 locates the author's analysis of these results as the gold standard. Single-round generation tends to blur the boundary between data description and analytical content.

G.6.4 Round 1 Prompt Template

```
You are an expert in empirical economics research. Read the attached
PDF paper, identify the paper's **core empirical/theoretical results**,
and extract a purely factual data description.

### Task 1: Scan all results in the paper

List all Tables and Figures in the paper (including titles). Label
the type and importance of each:
- Type: main_regression | robustness | descriptive | heterogeneity |
  mechanism | theory_result | other
- Importance: core | supporting | auxiliary

### Task 2: Select core results

Choose **1-2 results that best represent the paper's core
contribution**. Selection criteria:
- Prefer main regression tables (e.g., Table 1-3) or core theoretical-
  result figures
- Avoid robustness checks, descriptive statistics, and other auxiliary
  tables
- Selected results should allow substantive analysis

### Task 3: Write a rich context paragraph AND extract a purely
factual data description

**Context paragraph** (field: `context`, 150-250 words): For each
selected result, write a comprehensive background paragraph that
enables a reader who has NEVER seen the paper to understand the result.
This paragraph MUST include ALL of the following:
- **Research question & motivation** (2-3 sentences)
- **Data & sample** (1-2 sentences)
- **Methodology overview** (2-3 sentences)
- **What the result represents** (1 sentence)

**CRITICAL**: The context must NOT contain any interpretation,
mechanism discussion, or analytical conclusions.

**Data description** (field: `data_description`, 200-400 words):
Describe the data/numbers themselves in plain text:
- You may include: coefficient values, standard errors, significance
  levels, sample size, R-squared, direction of trends, etc.
- Do NOT include: causal interpretation, mechanism explanation, policy
  implications, links to theory

### Handling theoretical papers

If the paper is purely theoretical:
- `selected_results` should focus on core propositions, theorems, or
  comparative statics results
- `data_description` should state the formal result in plain language
  (what the theorem says) without explaining why it holds
- `result_type` should be `theoretical_result`
```

G.6.5 Round 2 Prompt Template

```
You are an expert in economics research methods. Re-read the attached
PDF paper and complete the following tasks.

### Task 1: Locate and extract a FOCUSED reference analysis

For each ‘selected_result‘ from Round 1, find the passage(s) in the
main text where the author **directly analyzes/interprets that result**.

**CRITICAL length constraint**: The reference_analysis must be
**400-600 words**. This means:
- Extract only the **most essential analytical paragraphs** that
  directly interpret the selected result
- Do NOT include extended discussions that cover other results, lengthy
  institutional background, or multi-page analyses
- If the author’s discussion spans multiple pages, **select and
  condense** to the 400-600 word core
- Prefer the paragraphs immediately following the result presentation

### Task 2: Tag applicable scoring dimensions

For each result, label which of the following 6 dimensions are
**actually supported by that result**. Use EXACTLY these dimension IDs:

- ‘accuracy‘: Can the result’s numbers/patterns be meaningfully checked
  for accuracy?
- ‘economic_interpretation‘: Does the result support substantive
  economic interpretation (magnitude, heterogeneity)?
- ‘mechanism‘: Does the result support discussion of economic mechanisms
  or causal channels?
- ‘theory_connection‘: Can the result be meaningfully connected to
  theoretical predictions or prior literature?
- ‘policy_implication‘: Does the result have implications for policy?
- ‘writing_quality‘: Always applicable -- assesses logical coherence
  and professional language

### Task 3: Build the complete task

Integrate data descriptions from Round 1 and reference analysis from
this round into the final evaluation item.
```

G.6.6 Evaluation Dimensions

Six dimensions are used consistently across Round 2 annotation and scoring prompts:

- **Accuracy** (`accuracy`): Correctness of numerical descriptions, trends, and statistical claims
- **Economic interpretation** (`economic_interpretation`): Substantive interpretation of magnitudes, heterogeneity, or practical significance
- **Mechanism** (`mechanism`): Explanation of economic mechanisms or causal logic behind results
- **Theory connection** (`theory_connection`): Linkage to theoretical predictions or prior literature
- **Policy implication** (`policy_implication`): Discussion of policy relevance (only when results support it)
- **Writing quality** (`writing_quality`): Logical clarity and professional language

Not every dimension applies to every result. `writing_quality` is always applicable; `accuracy` is almost always applicable. For theoretical results, `accuracy` refers to whether the formal statement is correctly characterized, and `mechanism` refers to the economic intuition behind the result.

G.6.7 Examinee Input

Examinees receive: (1) a context paragraph describing the research question, data, and methodology (150–250 words), and (2) a purely factual data description (e.g., "Table 3 shows that the coefficient on X is -0.37 (SE = 0.12), significant at the 1% level") without any analytical interpretation. They are asked to write a 300–500 word analysis paragraph interpreting the result across applicable dimensions.

G.6.8 Scoring Method

EA uses multi-dimension LLM-as-judge scoring (0–10 points per applicable dimension):

1. **Accuracy** (0–10): 9–10: all factual claims match data, no errors; 6–8: minor inaccuracies, core facts correct; 3–5: some factual errors affecting interpretation; 0–2: major factual errors.
2. **Economic interpretation** (0–10): 9–10: insightful interpretation with proper economic reasoning; 6–8: adequate interpretation, some depth lacking; 3–5: superficial, mostly restates numbers; 0–2: no meaningful interpretation.
3. **Mechanism** (0–10): 9–10: clear identification of driving mechanisms; 6–8: some mechanisms discussed but not fully developed; 3–5: vague or incomplete; 0–2: no mechanism identified.
4. **Theory connection** (0–10): 9–10: strong theoretical grounding with specific references; 6–8: some theoretical context; 3–5: weak or generic theory connection; 0–2: no theory connection.
5. **Policy implication** (0–10): 9–10: nuanced, well-grounded policy discussion; 6–8: reasonable policy points; 3–5: superficial; 0–2: no policy discussion.
6. **Writing quality** (0–10): 9–10: publication-ready prose, logical flow, precise language; 6–8: clear and organized, minor style issues; 3–5: readable but poorly organized; 0–2: disorganized or unclear.

Normalized score = total points / maximum possible points (counting only applicable dimensions, each worth 10 points max).

G.7 Summary

Table 10 summarizes the generation pipeline characteristics across all six tasks.

Task	Level	Rounds	Items/Paper	Scoring	Key Challenge
EK	L1	2	20 questions	Rule-based	Knowledge coverage balance
MM	L1	2	3 questions	LLM-as-Judge	Self-containment of questions
PA	L2	3	1 item	LLM-as-Judge	Clean abstract/body separation
LR	L2	4	1 item	Hybrid	Citation tier classification
MG	L2	3	1 item	LLM-as-Judge	Desymbolization fidelity
EA	L2	2	1–2 items	LLM-as-Judge	Data/analysis boundary

Table 10: Summary of task generation pipeline characteristics.

All six tasks employ multi-round generation pipelines with clear separation of concerns: early rounds perform exhaustive extraction or structural analysis with PDF access, while later rounds focus on question design, rubric construction, or validation using structured intermediate outputs. This design ensures high-quality, diverse evaluation items while maintaining consistency and reproducibility across the benchmark.

H Why AER-bench Omits a Human Expert Baseline

A natural question when presenting a benchmark like AER-bench is how large language models compare to human experts. After careful consideration, we have chosen not to collect a human performance baseline. This appendix details the reasons behind that decision.

Economic expertise is deeply fragmented. Modern economics is an enormously broad discipline, spanning macroeconomics, microeconomic theory, econometrics, labor, development, industrial organization, international trade, finance, and many other fields. A typical economics PhD student or even an experienced researcher usually possesses deep knowledge in only one or two of these areas. AER-bench draws tasks from 191 top-5 journal articles covering the full spectrum of contemporary economics. Consequently, no single human expert can be expected to perform competently across all seven task types and all subfields. Assembling a panel of specialists who could collectively cover the entire benchmark would not only be logistically daunting but would also introduce severe inter-annotator variability, making a single "human baseline" number difficult to interpret.

Human experts routinely rely on tools. In real-world research, economists seldom work from memory alone. They consult literature databases, use statistical software, and increasingly interact with large language models themselves. If we were to allow human participants to use such tools in order to simulate realistic research conditions, the resulting "human" baseline would in fact be a hybrid human–AI system. That would largely defeat the purpose: our aim is to measure how far pure LLMs are from performing end-to-end economic research, not to compare them against a human–AI combination that is not under controlled conditions.

Restricting human experts to a closed-book setting is unrealistic and difficult to enforce. Conversely, if we were to forbid any external aids, we would be testing an artificial skill—unaided recall of paper details—that is not how economists actually work. Moreover, ensuring that remotely participating experts do not quickly search for answers is practically impossible. The resulting data would be noisy and likely unrepresentative of genuine human expertise.

The benchmark already embeds a form of expert knowledge. It is worth noting that AER-bench already incorporates human expertise in two critical ways. First, all Concept Identification tasks are manually annotated by economics doctoral students, providing gold-standard labels. Second, the LLM-as-a-Judge rubrics are designed and weighted to reflect what a domain expert would value in an answer. These design choices ensure that the evaluation criteria are grounded in expert judgment, even without a separate human performance baseline.

Conclusion. Given these considerations—the breadth of expertise required, the confounding effect of tool use, the impracticality of pure closed-book testing, and the expert knowledge already embedded in the benchmark design—we believe that adding a human baseline would not strengthen the actionable insights that AER-bench provides. We instead focus on comparisons across models, which already reveal consistent performance hierarchies and systematic weaknesses. We encourage future work to explore targeted human–model comparisons on narrower subsets of the benchmark where domain specialists can be meaningfully recruited.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe AER-bench as a comprehensive benchmark for evaluating LLMs on economic research tasks, with claims supported by experimental results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated section, including LLM-as-a-Judge subjectivity, English-only scope, and temporal bias in historical concept identification.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper introduces a benchmark dataset and does not include theoretical results requiring formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 provides complete details on evaluation protocols, scoring methods, and model configurations. All rubrics are detailed in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The benchmark dataset, evaluation scripts, and complete documentation will be released under an open-source license upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 specifies all evaluation details including model APIs, temperature settings, judge model configurations, and scoring weight distributions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Each task contains approximately 200 items, providing sufficient sample size for statistically significant results. We report mean absolute differences between judges (2.6 pp) and task-level score distributions to demonstrate result robustness.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies that evaluations use commercial API endpoints (OpenAI, Anthropic, Google) with standard inference configurations, requiring no specialized compute infrastructure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics, using only publicly available academic papers and transparent evaluation methodologies.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses how AER-bench can guide responsible development of AI systems for economic research, while acknowledging risks of over-reliance on automated evaluation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The benchmark consists of curated academic papers and evaluation protocols, posing no misuse risks. No pre-trained models are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All source papers are from the American Economic Review, properly cited, and used in accordance with academic fair use principles for research purposes.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The AER-bench dataset includes comprehensive documentation of task definitions, rubrics, data sources, and evaluation protocols in the paper and supplemental materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The benchmark construction involved expert annotation by the authors, not crowdsourcing or human subjects research requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research. All data is derived from publicly available academic publications.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are the central focus of this benchmark. The paper evaluates 16 LLMs and uses LLM-as-a-Judge for scoring generative tasks, with full methodological details in Section 3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.