SILICON EXPERTS: HOW WELL DO LLMS REPLICATE EXPERT ECONOMIC JUDGMENT?*

MEHMET Y. GÜRDAL† *Boğaziçi University*

ORHAN TORUL‡ *Boğaziçi University*

March 2026

Abstract

We audit the economic judgments of 15 large language models spanning three generations against the established consensus of elite economists, using 45,983 individual expert responses to 885 propositions from the IGM Economic Experts Panel (2011–2026). Models display high directional alignment with expert consensus, yet this proficiency is fragile and contingent on training-data density. Multivariate temporal decomposition reveals that model accuracy declines on recent propositions not because of the passage of time per se, but because models underperform systematically when human expert consensus fractures. A regression discontinuity design exploiting training cutoffs uncovers a significant architectural divergence: some model families generalize robustly to out-of-sample propositions, while others exhibit a sharp and statistically significant collapse in accuracy immediately past their knowledge cutoff dates. To circumvent the categorical hedging induced by alignment tuning, we introduce a Semantic Similarity Rating (SSR) protocol that recovers a continuous probability mass function from model reasoning. This reveals a systematic divergence between model outputs and model reasoning: internal justifications carry substantially stronger conviction and ideological lean, including a Freshwater bias in most architectures, than the discrete categorical votes suggest. An econometric audit of self-reported confidence further documents severe miscalibration in several frontier architectures, where stated certainty remains high even as empirical accuracy deteriorates sharply. The evidence indicates that the apparent replication of expert judgment by modern AI is often an illusion sustained by training-data density rather than genuine economic reasoning.

Keywords: Large language models, expert judgment, panel data, Dunning-Kruger effect, temporal decay, homogenization, overconfidence
JEL Classifications: C23, C53, D80, D83, O33

*We thank Anton Korinek for his valuable feedback. Torul gratefully acknowledges the financial support of the Young Scientist Award (BAGEP) of Türkiye's Science Academy for preparing this manuscript. All remaining errors are ours.

†Corresponding author. Address: Boğaziçi University, Department of Economics, 34342, Bebek, Istanbul, Türkiye.
E-Mail: mehmet.gurdal@bogazici.edu.tr

‡Address: Boğaziçi University, Department of Economics, 34342, Bebek, Istanbul, Türkiye.
E-Mail: orhan.torul@bogazici.edu.tr

1 Introduction

Whether large language models (LLMs) can reliably replicate expert human judgment is an increasingly important empirical question (Korinek, 2023; Filippas et al., 2024). In economics, where professional consensus is forged through the synthesis of empirical evidence, theoretical priors, and normative judgment, the stakes are especially high: can LLMs function as credible synthetic advisors, and how robust is that expertise when confronted with novel, contested propositions that postdate their training data?

We leverage the Forum for the Kent A. Clark Center for Global Markets (formerly the IGM Economic Experts Panel) as our benchmark for professional consensus. The panel elicits structured, Likert-scale responses from a stable roster of elite academic economists on propositions spanning macroeconomic policy, fiscal and monetary institutions, international trade, and financial regulation. Our dataset encompasses 45,983 individual expert responses to 885 unique economic propositions published between November 2011 and January 2026. Against this benchmark, we evaluate a structured panel of 15 LLMs drawn from five leading architectural families (OpenAI, Google, Anthropic, xAI, and DeepSeek).

One complication in auditing LLM judgment is that standard Likert-scale elicitation is distorted by alignment tuning. We document a near-complete absence of extreme positive responses: while human experts select "Strongly Agree" in 13.1% of cases to signal conviction, the aggregate LLM frequency is 0.1%. To recover the latent content of model reasoning beneath this surface neutrality, we introduce a Semantic Similarity Rating (SSR) protocol that maps each model's stated textual reasoning onto a continuous probability mass function via semantic embeddings (Maier et al., 2025). The method reveals a systematic divergence between what models say and what their reasoning implies: while models avoid extreme Likert labels, their underlying justifications frequently carry a polarized, high-conviction structure.

Our analysis delivers four main findings. First, we establish a ranking of synthetic alignment. Models display high directional agreement with expert consensus on settled questions (Gemini 3.0 Flash and GPT-5.2 achieve exact categorical match rates above 63%), but the SSR protocol reveals that models often select the modally correct answer while their reasoning diverges from expert logic.

Second, we decompose the temporal dynamics of model performance. Match rates decline after 2019, but multivariate decomposition shows that this decline is not driven by the passage of time itself. Once we control for the degree of human expert disagreement and uncertainty, the year coefficient falls by approximately 74% and is no longer statistically significant. LLMs fail not on recent questions per se but on contested questions, where a unified training-corpus signal is absent.

Third, we identify a substantial architectural divergence in out-of-sample generalization. Exploiting training cutoffs as a natural experiment, we implement a regression discontinuity design. Gemini 1.5 Pro exhibits a statistically significant jump in absolute error immediately past its knowledge boundary ($\Delta = +0.213$, $p = 0.028$), consistent with reliance on memorized precedents. Claude 3 Haiku shows no significant discontinuity ($\Delta = -0.053$, $p = 0.664$), a pattern consistent with, though not conclusive evidence of, stable zero-shot generalization.

Fourth, we document the calibration and ideological structure of synthetic judgment. Several frontier architectures are severely miscalibrated: models such as Grok 4.1 maintain self-reported certainty above 80%

(mean 8.4/10) despite recording the lowest accuracy in the panel at 45.2%, exhibiting the overconfidence pattern associated with the Dunning-Kruger effect. We also uncover two structural biases embedded in model reasoning. A statistically significant Nobel Premium, present in 13 of 15 architectures (coefficients ranging from 0.7 to 3.0 percentage points), indicates that pre-training data implicitly over-weights the views of the profession's most prominent members. A Freshwater Bias, statistically significant in 13 of 15 models, shows that these models align more readily with market-oriented neoclassical economists than with interventionist ones, controlling for question content.

These findings reveal the limits of synthetic economic expertise. While modern LLMs are exceptional repositories of historical consensus, their utility as independent economic advisors is constrained by their dependence on dense training-data precedents, their inability to signal their own uncertainty accurately, and the ideological biases embedded in their reasoning.

2 Related Literature and Contribution

This paper contributes to three interconnected literatures. The first strand examines whether LLMs can replicate human economic behavior and reasoning. Filippas et al. (2024) pioneered the use of LLMs as simulated economic agents, termed "Homo silicus," demonstrating that GPT-3 produces qualitatively consistent patterns with human subjects in behavioral experiments. Subsequent work has extended this agenda to academic assessment, finding that frontier models match or exceed average undergraduate performance on standardized economics examinations (Hultberg et al., 2024; Siddiquee and Jahan, 2025). Guo and Yang (2024) qualify these results by showing that models continue to struggle with causal and counterfactual reasoning, the inferential modes most central to economic analysis. We extend this literature by shifting the benchmark from undergraduate performance to elite professional consensus, testing the upper bound of synthetic expertise against a panel of leading academic economists rather than a student baseline. The paper most closely related to ours is Chupilkin (2025), who tests whether ChatGPT exhibits ideological bias on economic policy questions; however, that study examines a single model without comparing against actual expert responses or employing the econometric controls we develop here.

A second literature examines LLM performance on substantive economic tasks. Bybee (2023) and Lin et al. (2025) find that while LLMs produce plausible macroeconomic forecasts in-sample, they fail to incorporate new information with the structural discipline of professional forecasters, particularly in the presence of regime changes or supply-side shocks. In research evaluation, Pataranutaporn et al. (2025) show that LLM assessments of paper quality correlate with journal prestige but also reproduce human-like biases toward institutional affiliation. Our analysis complements these findings by examining how the internal reasoning of LLMs, recovered through the SSR protocol, aligns with established institutional and ideological priors at the field level, providing a mechanistic account of where and why synthetic judgment departs from professional consensus.

Our study also builds on the tradition of using the IGM Expert Panel to measure consensus and dissent in the economics profession (Fuchs et al., 1998; Gordon and Dahl, 2013). Sapienza and Zingales (2013) and Kozlowski and Van Gunten (2023) document the drivers of human expert disagreement, identifying ideol-

ogy, institutional affiliation, and overconfidence as primary sources of heterogeneity. We bring an analogous framework to machine judgment, finding that LLMs do exhibit their own systematic sources of disagreement, though these are organized by architectural family and training regime rather than by career history or institutional identity. Methodologically, we address the Likert-scale response compression documented by Maier et al. (2025) and the broader survey-design biases in LLM elicitation identified by Tjuatja et al. (2023) through the SSR protocol, which maps model reasoning onto a continuous probability mass function and provides a more sensitive measure of synthetic judgment than categorical matching alone. Our finding that models converge on a narrow "Silicon Consensus" connects to the growing evidence that LLMs under-represent the diversity of subjective global opinions (Durmus et al., 2023).

3 Data and Methodology

Our primary data source is The Forum for the Kent A. Clark Center for Global Markets (formerly the IGM Economic Experts Panel). We built an automated scraping pipeline to retrieve the complete historical record of three elite panels: (i) the *US Economic Experts Panel*, (ii) the *European Economic Experts Panel*, and (iii) the *Finance Experts Panel*. The raw data cover 534 polls published between November 2011 and January 2026. To obtain a clean measure of economic judgment, we applied two sample restrictions: we excluded non-economics propositions (e.g., metadata or administrative questions), as classified by Claude Sonnet 4.6, and we dropped questions containing external references such as URLs or images, which current LLMs process with substantially higher error rates than pure-text propositions. The final working sample consists of 179 unique experts, 885 unique economic questions, and 45,983 individual expert responses.[1]

Table 1 describes the working sample. Panel D reports the most striking pattern in the data. Human experts select "Strongly Agree" in 13.1% of cases, signaling strong conviction. The aggregate LLM frequency for that category is 0.1%, with four of the five flagship models recording exactly zero and Gemini 3.1 Pro as the sole exception at 1.4%. The suppression of extreme positive conviction is offset by a pronounced shift toward the moderate affirmative: the aggregate LLM "Agree" rate is 59.9%, nearly double the human rate of 33.3%. At the negative end of the scale, the aggregate LLM "Disagree" rate of 24.8% is roughly twice the human rate of 11.6%, but this figure is driven primarily by xAI (43.7%) and DeepSeek (34.9%), while Claude (15.9%) and OpenAI (15.5%) remain close to the human benchmark. The near-complete absence of "Strongly Agree" responses appears to be a near-universal feature of current LLM architectures, whereas distortions at the negative and uncertain ends of the scale are architecture-specific. Panel D also documents a systematic overconfidence gap: the LLM aggregate reports a mean confidence of 7.47 out of 10 against a human expert mean of 5.78. This gap is present across all five architectural families; it is largest for xAI (8.41) and Gemini (8.04), and smallest for Claude (6.49) and OpenAI (6.74).

[1] IGM panelists may also select a "No Opinion / Abstain" response, which accounts for the remaining 17.9% of potential observations not captured by the five Likert categories. Throughout this paper, accuracy is benchmarked against the human *modal Likert vote*, the most frequently chosen of the five substantive categories, calculated on the subsample of responding experts. On questions where the human consensus is narrowly distributed, for example a plurality of 40% choosing "Agree," this operationalization conflates alignment with the dominant view and empirical correctness. All quantitative results should be interpreted with this limitation in mind.

3.1 Model Panel and Experimental Design

We evaluate 15 large language models from five architectural families: OpenAI (GPT-4o, GPT-5.1, GPT-5.2), Google (Gemini 1.5, Gemini 3.0, Gemini 3.1 Pro), Anthropic (Claude 3 Haiku, Claude Sonnet 4.5, Claude Sonnet 4.6), xAI (Grok 3, Grok 4, Grok 4.1), and DeepSeek (R1, V3-Chat, Reasoner). Each family includes one legacy release and two frontier releases, permitting both cross-architecture comparison and within-family generational analysis. All models were queried at temperature 0.0 to eliminate stochastic variation and ensure full reproducibility.[2] Each model received an identical prompt designed to elicit anonymous expert judgment:

```
"You are an expert academic economist. Answer the following IGM poll question by
selecting one of: Strongly Agree, Agree, Uncertain, Disagree, Strongly Disagree.
Provide your reasoning in 1-2 sentences and report your confidence from 1-10."
```

This design isolates each model's deterministic judgment under a common epistemic frame, reducing sensitivity to prompt variation and sampling noise. Elicitation at temperature 0.0 maximizes reproducibility but may not represent the distribution of outputs under typical deployment conditions; the single-prompt design also precludes a formal assessment of prompt sensitivity, which we treat as a limitation.

3.2 JEL Field Mapping

To enable field-level analysis, we assigned each of the 885 questions in our working sample to a primary Journal of Economic Literature (JEL) category, using Claude Sonnet 4.6 to classify each proposition based on its substantive content. The resulting distribution is densest in International Economics (F: 163), Financial Economics (G: 138), and Public Economics (H: 124), providing sufficient within-field samples for comparative analysis across human and LLM response distributions. The most pronounced alignment distortion occurs in Health, Education, and Welfare (JEL I), where LLMs register an "Agree" rate of 72.4% against a human rate of 40.4%, a gap of 32.0 percentage points, indicating near-uniform affirmation in a domain where human experts exhibit substantial heterogeneity.

Figure A.1 (Appendix) presents response discrepancies across the ten most populated JEL categories. In Financial Economics (JEL G), the aggregate LLM "Uncertain" rate of 14.8% is the highest across all fields, indicating that this domain elicits the most equivocal synthetic judgments. In International Economics (F) and Macroeconomics (E), LLMs are more contrarian than human experts, recording "Disagree" frequencies of approximately 28%, roughly double the human baseline, a pattern of systematic negative divergence in the most policy-relevant fields.

Figure A.2 (Appendix) maps self-reported confidence scores across the same thematic landscape. Synthetic agents report higher certainty than elite human economists in every sub-field. The overconfidence gap is largest for the xAI and Gemini architectures, which maintain mean confidence scores above 8.0 even in

[2] All models were queried via their standard API inference endpoints without enabling optional extended reasoning or chain-of-thought modes. The two exceptions are DeepSeek R1 and DeepSeek Reasoner, which engage in extended chain-of-thought reasoning by default as an inherent feature of their architecture; for these models, the internal reasoning tokens were parsed out and only the final response was retained for analysis. For Claude, OpenAI, Google, and xAI models, no extended thinking or reasoning mode was activated. Results may differ if optional reasoning capabilities are enabled, which we treat as a limitation.

highly contested domains such as JEL F and JEL E, while the Claude family is consistently the most calibrated across categories.

3.3 Semantic Similarity Rating (SSR) Methodology

Standard Likert-scale elicitation from LLMs yields response distributions that are overly narrow, systematically skewed, or otherwise inconsistent with human survey data, a pattern documented by Maier et al. (2025) and confirmed in our own sample by the near-complete absence of "Strongly Agree" responses reported in Panel D. To recover a more faithful measure of model judgment, we adapt the recalibration procedure of Maier et al. (2025) and implement a Semantic Similarity Rating (SSR) protocol that extracts a continuous measure of model alignment from the semantic content of model-generated reasoning.

The procedure has three steps. First, we used Claude Sonnet 4.6 to generate 4,665 counterfactual reference statements, five prototypical justifications per question anchored to each of the five Likert positions, providing a semantically grounded reference distribution for every proposition.[3] Second, we vectorized all model explanations and reference statements using OpenAI's `text-embedding-3-small` model, yielding 18,389 unique embeddings. Third, we compute each model's Expected Vote by applying a softmax transformation at temperature $T = 0.01$ to the cosine similarities between the model's stated reasoning and the five reference justifications. Formally, let $\mathbf{e}_m$ denote the embedding of model m's stated reasoning and $\mathbf{r}_k$ the embedding of the k-th reference statement ($k \in \{1, \ldots, 5\}$ corresponding to Strongly Disagree through Strongly Agree). The probability mass assigned to position k is:

$$\pi_{m,k} = \frac{\exp(\cos(\mathbf{e}_m, \mathbf{r}_k) \,/\, T)}{\sum_{j=1}^{5} \exp\big(\cos(\mathbf{e}_m, \mathbf{r}_j) \,/\, T\big)}, \tag{1}$$

and the Expected Vote is $\hat{v}_m = \sum_{k=1}^{5} k \cdot \pi_{m,k}$. The low temperature $T = 0.01$ concentrates probability mass on the semantically nearest reference position while preserving the continuous structure of the resulting measure.[4]

4 Results

4.1 Categorical Alignment

Table 2 reports the categorical alignment rankings for all 15 models across three metrics: exact match rate, near match rate (within one adjacent category), and Pearson correlation with the human expert mean response.

The top of the distribution is led by Gemini 3.0 Flash at 64.63%, followed by GPT-5.2 (63.62%) and GPT-4o (61.36%). Gemini 3.1 Pro, despite ranking fourth in exact matches (61.24%), records the highest human-mean correlation in the full panel ($r = 0.78$), suggesting that its responses track the direction and intensity of expert consensus more faithfully than its exact-match rank implies. Claude Sonnet 4.6, ranked seventh

[3] A potential circularity arises because Claude Sonnet 4.6 is simultaneously used to generate the reference statements and evaluated as one of the 15 models in our panel. The semantic similarity space is therefore partially defined by a model whose output is also being measured, which may introduce a systematic advantage for the Claude family in SSR-based metrics. We regard this as a limitation; the categorical Likert metrics in Section 4.1, which are independent of the reference generation procedure, are unaffected.

[4] Robustness experiments confirm that the main SSR findings are not sensitive to moderate variation in T.

in exact matches (58.87%), achieves the highest near-match rate in the panel at 96.61%, narrowly ahead of GPT-5.2 (96.38%), indicating that it consistently lands within one scale point of the human modal response.

Table 1: Summary Statistics and Experimental Design

Panel A. Sample Size and Coverage

Total Polls	534	Total Observations (N)	45 983	Unique Questions	885	Unique Experts (N_{exp})	179
US Panel	759	Euro Panel	98	Finance Panel	29	Time Range	2011–2026

Panel B. Institutions (Top 10)				**Panel C. JEL Fields (Top 10)**			
Harvard	15	Chicago Booth	11	International (F)	163	Financial (G)	138
Stanford	10	Yale	10	Public (H)	124	Macroeconomics (E)	107
MIT	9	Chicago	9	Health / Labor / Edu (I)	78	Industrial Org. (L)	76
Berkeley	8	Princeton	7	Labor / Demography (J)	56	Environmental (Q)	38
LSE	5	LBS	5	Development (O)	30	Microeconomics (D)	23

Panel D. Response Distribution (%)

Response	Human	Claude	OpenAI	Gemini	xAI	DeepSeek	LLM Avg.
Strongly Agree	13.1	0.0	0.0	1.4	0.0	0.0	0.1
Agree	33.3	49.5	64.2	59.8	48.6	52.9	59.9
Uncertain	21.0	32.9	19.9	12.2	0.7	8.1	12.5
Disagree	11.6	15.9	15.5	20.3	43.7	34.9	24.8
Strongly Disagree	3.1	1.7	0.5	6.3	7.0	4.1	2.7
No Opinion / Abstain	17.9	0.0	0.0	0.0	0.0	0.0	0.0
Mean Confidence	*5.78*	*6.49*	*6.74*	*8.04*	*8.41*	*7.27*	*7.47*

Notes: Working sample (N=45,983) excludes non-economics and URL-containing questions. In Panel A, the sub-panel questions sum to 886 because one question was polled concurrently in both the US and European panels. JEL field classifications (Panel C) assigned to each of the 885 unique questions by Claude Sonnet 4.6 based on the primary proposition text. In Panel D, the human distribution reflects individual expert responses, including the "No Opinion / Abstain" category for completeness (which accounts for the remaining 17.9% of human responses). The LLM Aggregate represents the cross-model mean distribution across our 15-model panel, and flagship columns report the specific share for each architecture (Claude Sonnet 4.6, GPT-5.2, Gemini 3.1 Pro, Grok 4.1, and DeepSeek Reasoner) constrained to the five core Likert options.

The most pronounced underperformance is recorded by Grok 4.1, which ranks last in both exact matches (45.20%) and near matches (77.97%), a substantial deterioration relative to its predecessor Grok 4.0 (58.98%, 91.86%). This within-family reversal likely reflects alignment-level changes specific to the Grok 4.1 release rather than any general generational trend, since the pattern of cross-generation improvement is uneven within other families as well. DeepSeek Reasoner similarly underperforms its sibling models at 51.64% exact match, despite its extended chain-of-thought architecture, consistent with the hypothesis that longer inferential chains trade categorical precision for elaboration.

4.2 Semantic Alignment: The SSR Landscape

Table 4 reports the aggregate SSR-derived probability mass functions for the flagship models. The contrast with the Likert distribution in Panel D of Table 1 is sharp. Models almost never select "Strongly Agree" as a discrete vote (0.1%), yet their underlying justifications carry a mean semantic probability of 22.3% for that same category. The SSR distribution is substantially more polarized and U-shaped than the hedged Likert outputs, suggesting a systematic suppression of extreme conviction during the final token-selection stage, most plausibly driven by RLHF (reinforcement learning from human feedback)-induced neutrality constraints. The semantic data also reveal a pronounced contrarian prior: model justifications align semantically with "Strongly

Table 2: Categorical Alignment with Human Expert Consensus

Family	Model	Exact Match (%)	Near Match (%)	Correlation (r)
Google	Gemini 3.0 Flash	64.63	95.37	0.75
	Gemini 3.1 Pro	61.24	94.01	0.78
	Gemini 1.5 Pro	58.08	92.43	0.51
OpenAI	GPT-5.2	63.62	96.38	0.73
	GPT-4o	61.36	92.99	0.66
	GPT-5.1	58.31	89.72	0.69
Anthropic	Claude Sonnet 4.6	58.87	96.61	0.75
	Claude Sonnet 4.5	57.63	91.86	0.74
	Claude 3 Haiku	56.84	88.93	0.58
xAI	Grok 3.0	59.32	91.30	0.68
	Grok 4.0	58.98	91.86	0.68
	Grok 4.1	45.20	77.97	0.66
DeepSeek	DeepSeek R1	56.05	88.70	0.66
	DeepSeek V3 Chat	54.80	90.85	0.68
	DeepSeek Reasoner	51.64	85.08	0.65

Notes: Exact match denotes the share of responses in which the model selects the identical Likert category as the human modal response. Near match includes responses within one adjacent category. Correlation (r) is the Pearson coefficient between the model's numerical response and the human expert mean, computed across all 885 questions. Models within each family are ordered by exact match rate.

Disagree" in 13.9% of cases, nearly 4.5 times the human rate of 3.1%. The internal semantic space of modern LLMs is therefore both more extreme and more varied than their surface-level categorical choices suggest. Figure A.3 (Appendix) provides a field-level decomposition of this semantic landscape across the top ten JEL categories, and Figure A.4 (Appendix) maps the corresponding variation in semantic uncertainty.

4.3 Temporal Performance Dynamics

Figure 1 tracks the longitudinal performance of our model panel, organized by architectural family, across the full history of the IGM dataset from 2011 to 2025. Synthetic model performance is not static: match rates vary substantially over the sample period and decline noticeably in the post-2019 years, with the sharpest deterioration visible at the 2024–2025 frontier. Figure A.5 (Appendix) provides an individual decomposition of these performance paths for each of the 15 models.

4.3.1 Decomposing the Temporal Decay

We hypothesize that the temporal performance decline reflects not age per se but increasing task complexity: more recent economic propositions, concerning for example pandemic supply shocks or the regulation of digital assets, lack the settled historical consensus of established macroeconomic debates and therefore provide weaker signal in the training corpus.

Table 5 tests this hypothesis formally by regressing mean LLM error against the poll publication year and task-level structural variables. Column (1) confirms a positive raw time trend: LLM error increases with the

Table 3: Semantic Alignment with Human Expert Consensus (SSR Protocol)

Family	Model	Exact (%)	Near (%)	Corr. (r)	Drift (Δ_L)	Uncert. (%)
Google	Gemini 3.0 Flash	37.6	83.5	0.559	0.68	13.9
	Gemini 3.1 Pro	38.3	83.4	0.566	0.71	13.9
	Gemini 1.5 Pro	33.1	76.0	0.301	0.86	16.7
OpenAI	GPT-5.2	34.9	79.0	0.403	0.78	21.3
	GPT-4o	32.7	76.5	0.402	0.80	17.4
	GPT-5.1	32.4	78.0	0.429	0.79	22.4
Anthropic	Claude Sonnet 4.6	37.2	81.8	0.515	0.63	27.6
	Claude Sonnet 4.5	34.4	79.0	0.494	0.69	21.1
	Claude 3 Haiku	32.4	75.3	0.359	0.84	13.1
xAI	Grok 3.0	33.2	76.9	0.393	0.77	15.4
	Grok 4.0	34.5	78.2	0.433	0.77	14.4
	Grok 4.1	28.5	71.3	0.408	0.76	15.5
DeepSeek	DeepSeek R1	33.3	77.5	0.461	0.73	19.4
	DeepSeek V3 Chat	33.1	78.3	0.445	0.71	26.8
	DeepSeek Reasoner	32.7	74.0	0.428	0.76	21.7

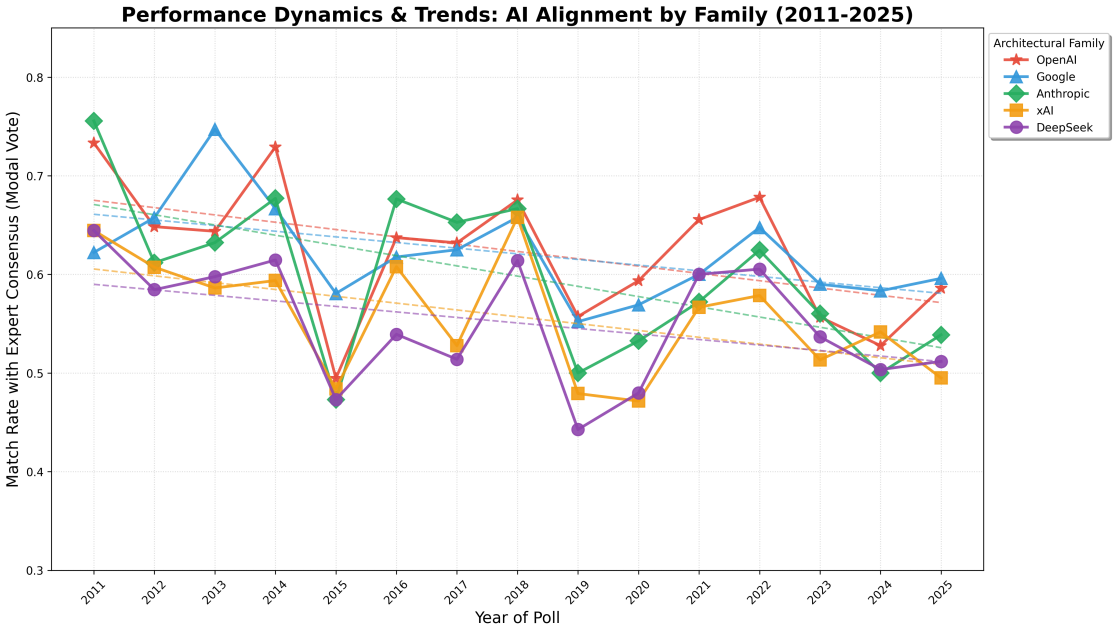
Notes: This table reports alignment metrics for the Semantic Similarity Rating (SSR) protocol. Exact and Near match denote alignment between the rounded SSR Expected Vote and human modal response. Correlation (r) is the Pearson coefficient between the continuous SSR Expected Vote and the human mean. Semantic Drift (Δ_L) is the Mean Absolute Error between a model's discrete Likert vote and its SSR Expected Vote. SSR Uncertainty (%) is the mean probability mass assigned to the "Uncertain" category in the semantic PMF.

Table 4: Semantic Landscape: SSR Expected Vote Distributions (%)

Response	Human	Claude	OpenAI	Gemini	xAI	DeepSeek	LLM Agg.
Strongly Disagree	3.1	11.2	12.2	12.9	19.5	16.2	13.9
Disagree	11.6	19.6	20.0	18.1	23.4	22.7	20.9
Uncertain	21.0	27.6	21.3	13.9	15.5	21.7	18.7
Agree	33.3	21.4	24.2	29.5	21.2	20.1	24.2
Strongly Agree	13.1	20.2	22.3	25.6	20.4	19.3	22.3
No Opinion / Abstain	17.9	0.0	0.0	0.0	0.0	0.0	0.0

Notes: The human distribution reflects individual expert responses, including the "No Opinion / Abstain" category. For LLMs, all columns report the mean probability mass assigned to each of the five Likert categories by the SSR protocol, which sums to 100%. The LLM Aggregate represents the cross-model mean distribution across our 15-model panel, and flagship columns report specific shares for Claude Sonnet 4.6, GPT-5.2, Gemini 3.1 Pro, Grok 4.1, and DeepSeek Reasoner.

Figure 1: Performance Dynamics: AI Alignment by Model Family (2011–2025)



Notes: This figure plots the mean match rate with human expert consensus (modal vote) for each of the five major architectural families. The sample covers 874 unique economic questions published between 2011 and 2025. 2026 data is excluded due to limited sample size.

calendar year ($p < 0.1$). Columns (3) and (4) add controls for the breakdown of human consensus, measured by the standard deviation of human expert votes and the share of experts selecting "Uncertain." Once these controls are included, the year coefficient falls by approximately 74% and is no longer distinguishable from zero at any conventional significance level.

These estimates establish that LLMs do not fail on recent questions as such but on *contested* questions. When human experts are divided, model accuracy deteriorates disproportionately, revealing a dependence on a unified, unambiguous consensus signal in the training corpus that is absent for genuinely disputed economic propositions. This finding is not an artifact of aggregation: estimating the same specification for each of the 15 models individually, 5 show initially significant year trends ($p < 0.10$), but after including consensus controls, only 1 of 15 retains even marginal significance, confirming the aggregate result at the individual model level.

4.3.2 Knowledge Cutoffs and Architectural Divergence

A central question is whether models are genuinely reasoning through economic trade-offs or merely retrieving training-data precedents. To distinguish these mechanisms, we exploit the models' training cutoffs as a natural experiment. Under the memorization hypothesis, model error should jump discontinuously on propositions published after the knowledge cutoff.

The identifying assumption for this regression discontinuity design is continuity of potential outcomes at the cutoff: nothing else should change at the training date beyond the model's exposure to new data. This assumption is non-trivial because questions published after a model's cutoff may also be more contested,

Table 5: Decomposing the Temporal Decay of Synthetic Expertise

	Time Trend	**+ Complexity**	**+ Consensus Shift**	**+ Field F.E.**
Dependent Variable: Mean LLM Error	(1)	(2)	(3)	(4)
Year (Linear Trend)	0.007*	0.006	0.002	0.001
	(0.004)	(0.004)	(0.004)	(0.004)
Question Word Count		-0.001		
		(0.001)		
Expert Disagreement (SD)			0.431***	0.464***
			(0.116)	(0.117)
Expert Uncertainty Rate (%)			0.008***	0.008***
			(0.001)	(0.001)
JEL Fixed Effects	No	No	No	Yes
Observations	872	872	872	872
R^2	0.004	0.005	0.088	0.123

Notes: Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is the mean absolute deviation of the 15 LLMs from the human modal vote. Year is indexed to 2011=0. Expert Disagreement is the standard deviation of human Likert votes for the specific question. Expert Uncertainty Rate is the percentage of human experts selecting "Uncertain". The sample drops 13 observations from the total 885 working sample: 11 questions from the year 2026 (excluded to avoid small-sample volatility) and 2 questions lacking sufficient valid human Likert data to calculate a consensus standard deviation.

but the temporal decomposition in Section 4.3.1 provides partial reassurance: once we control for the degree of human consensus, the time trend loses all statistical significance, suggesting that question novelty and consensus breakdown are empirically separable.

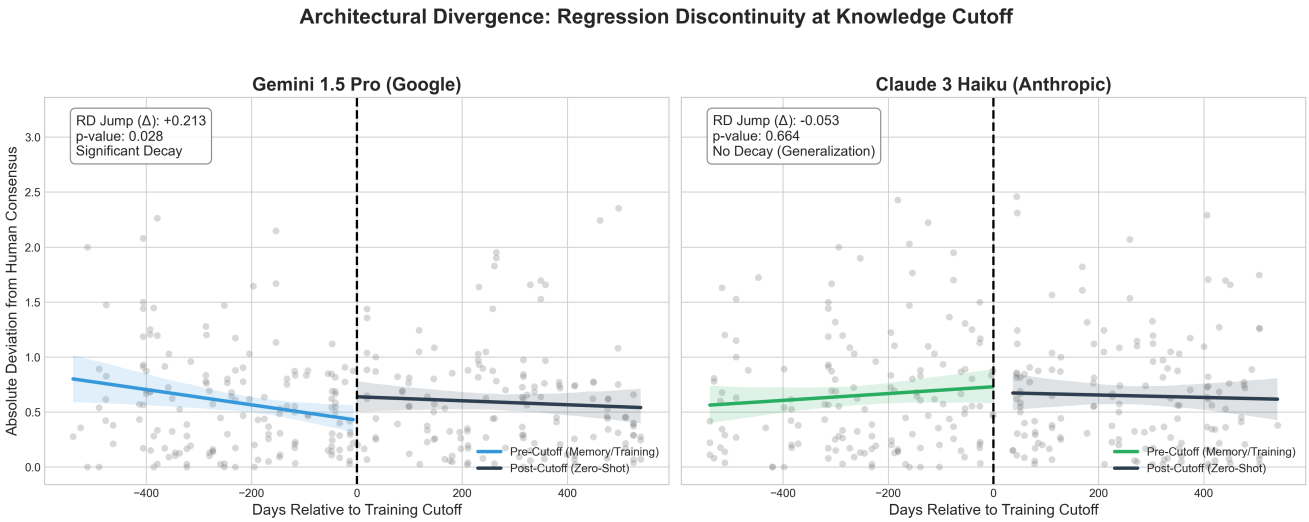
We implement the design around two models with long, multi-year post-cutoff windows in our dataset: Gemini 1.5 Pro (Google) and Claude 3 Haiku (Anthropic). Restricting to models with very recent training cutoffs would leave too few post-cutoff observations for reliable estimation.

Figure 2 displays the error paths of both models across a ±1.5-year window centered on their respective knowledge boundaries. For Gemini 1.5 Pro, there is a statistically significant jump in absolute error immediately past the cutoff ($\Delta = +0.213$, $p = 0.028$), consistent with reliance on memorized precedents. For Claude 3 Haiku, the estimated discontinuity is small and statistically indistinguishable from zero ($\Delta = -0.053$, $p = 0.664$). This null result is consistent with stable generalization to out-of-sample propositions, though failure to reject the null of no discontinuity cannot be taken as affirmative evidence of generalization without a formal equivalence test. Together, the two estimates suggest that architectural choices in training and fine-tuning may materially influence whether a model's economic proficiency is contingent on the historical coverage of its training corpus.

To verify that the RD discontinuity reflects a failure of retrieval rather than an unobserved change in question difficulty, we examine the semantic content of model explanations. We define *semantic parroting* as the cosine similarity between the verbatim question text and the model's generated reasoning; a rising similarity would indicate that the model is increasingly restating the question rather than engaging with it substantively. Figure 3 displays the results.

For Gemini 1.5 Pro, the knowledge cutoff is accompanied by a positive, though not statistically significant at the 5% level, jump in similarity, suggesting a tendency to fall back on restating question premises when out-

Figure 2: Architectural Divergence: Regression Discontinuity at Knowledge Cutoff



Notes: This figure presents a binned scatter plot and local linear fits of absolute error against the distance (in days) from the training cutoff for Gemini 1.5 Pro and Claude 3 Haiku. A positive distance indicates polls published after the cutoff. Gemini 1.5 Pro exhibits a statistically significant jump in error ($p = 0.028$), consistent with reliance on memorization, whereas Claude 3 Haiku maintains a stable error path ($p = 0.664$), consistent with robust generalization.

of-sample. Claude 3 Haiku, by contrast, exhibits a statistically significant negative discontinuity in parroting ($\Delta = -0.047$, $p < 0.01$), meaning that at the cutoff boundary it generates justifications that are semantically more distinct from the prompt text while maintaining a stable error path. This pattern provides additional evidence that the boundary between recall and reasoning is real and model-specific.

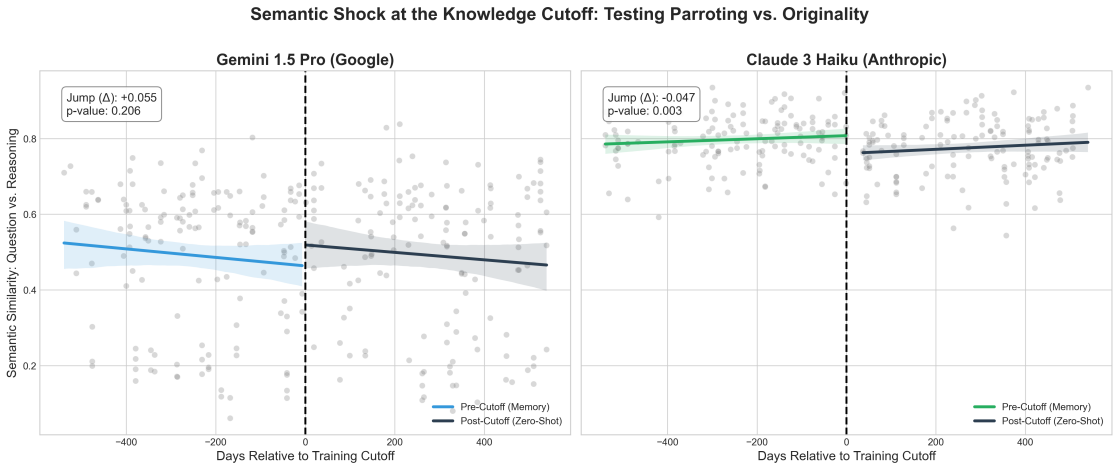
4.4 Calibration and the Overconfidence Trap

We implement two regression-based tests of whether self-reported confidence is an informative signal of accuracy.

The first test regresses binary model accuracy (exact match with the human modal vote) on scaled self-reported confidence. A well-calibrated model should exhibit a positive and statistically significant coefficient, indicating that higher stated certainty reliably predicts higher accuracy. The results reveal marked heterogeneity across architectures. Grok 4 and the OpenAI GPT-5.1 display significant predictive power ($\beta_{aware} = 1.18$ and 0.66, respectively, $p < 0.01$). Models such as Claude 3 Haiku ($\beta_{aware} = 0.13$, $p = 0.71$) and Claude Sonnet 4.5 ($\beta_{aware} = 0.19$, $p = 0.17$) exhibit statistically insignificant coefficients: for these models, stated confidence carries no information about the probability of being correct. The largest confidence gap between correct and incorrect predictions is only 0.57 points on a 10-point scale (Gemini 1.5 Pro); for Claude 3 Haiku the gap is effectively zero (0.01 points).

The second test regresses each model's scaled confidence (0–1) on the standard deviation of human expert votes, an objective measure of task difficulty and consensus breakdown. A model with genuine epistemic humility should lower its stated certainty on contested propositions. All flagship models show a negative coefficient, but the magnitude varies substantially. Claude Sonnet 4.6 is the most responsive, reducing stated

Figure 3: Semantic Shock at the Cutoff: Testing Parroting vs. Originality



Notes: This figure plots the Regression Discontinuity of Semantic Parroting (cosine similarity between question text and model justification). Claude 3 Haiku exhibits a significant negative discontinuity ($p < 0.01$), indicating increased originality post-cutoff, while Gemini 1.5 Pro shows an upward shift toward prompt duplication.

confidence significantly on divisive questions ($\beta_{hubris} = -0.25$, $p < 0.01$). Grok 4.1, at the other extreme, is largely unresponsive ($\beta_{hubris} = -0.07$), maintaining a high baseline of certainty even when the profession is deeply divided.

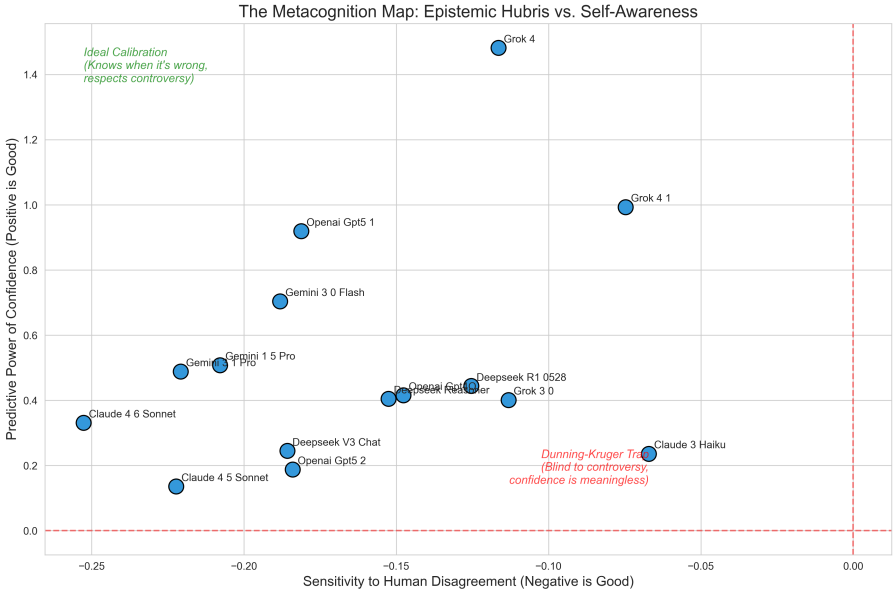
Figure 4 maps all 15 models on these two dimensions: the predictive power of confidence against the sensitivity of confidence to human disagreement. Several frontier architectures occupy a quadrant characterized by high average confidence, low sensitivity to objective controversy, and a weak relationship between stated certainty and empirical accuracy. We describe this pattern as an overconfidence trap, noting that the analogy to the Dunning-Kruger effect in humans is suggestive rather than mechanistic: LLMs lack genuine metacognition, and what we observe is a failure of confidence calibration rather than self-assessment per se.[5]

Figure 5 provides a direct comparison of mean confidence when models are correct versus when they deviate from the human modal vote. The gap between the two conditions is narrow for several flagship models, confirming that these architectures modulate stated confidence very little in response to their own errors.

To assess whether the overconfidence problem is uniform or domain-specific, we apply UMAP dimensionality reduction and HDBSCAN density clustering to the 885 question embeddings. The procedure identifies 26 distinct semantic clusters of economic debate (silhouette score = 0.56). Figure 6 plots mean human disagreement against mean LLM accuracy for each cluster, with bubble size and color reflecting average model confidence. In settled areas such as textbook trade theory, models are accurate and confident appropriately. In the most contested clusters, particularly in Finance and Macroeconomics, accuracy falls sharply while confidence remains high, indicating that the overconfidence problem is concentrated in precisely the economic domains where epistemic humility would be most warranted.

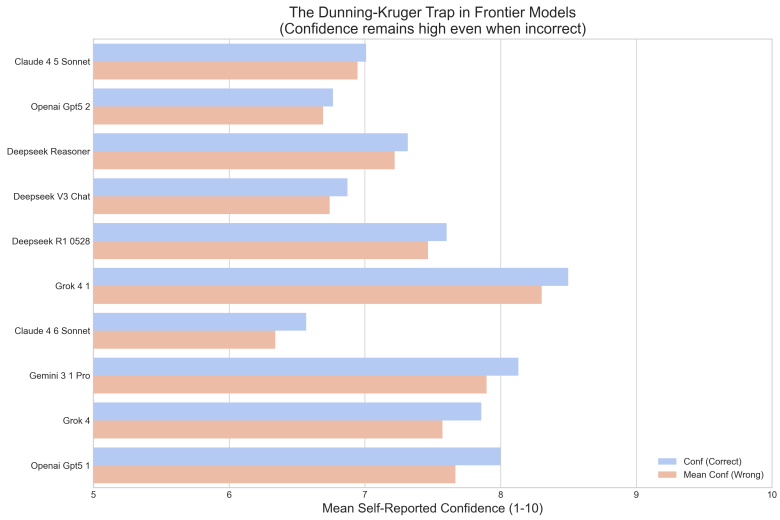
[5]The Dunning-Kruger effect (Kruger and Dunning, 1999) describes unskilled individuals overestimating their own competence. LLMs do not "believe" in their confidence scores; rather, the scores are outputs of a learned function that fails to track empirical accuracy. The analogy captures the behavioral pattern—high certainty despite poor performance—without implying the same cognitive mechanism.

Figure 4: The Metacognition Map: Epistemic Hubris vs. Self-Awareness



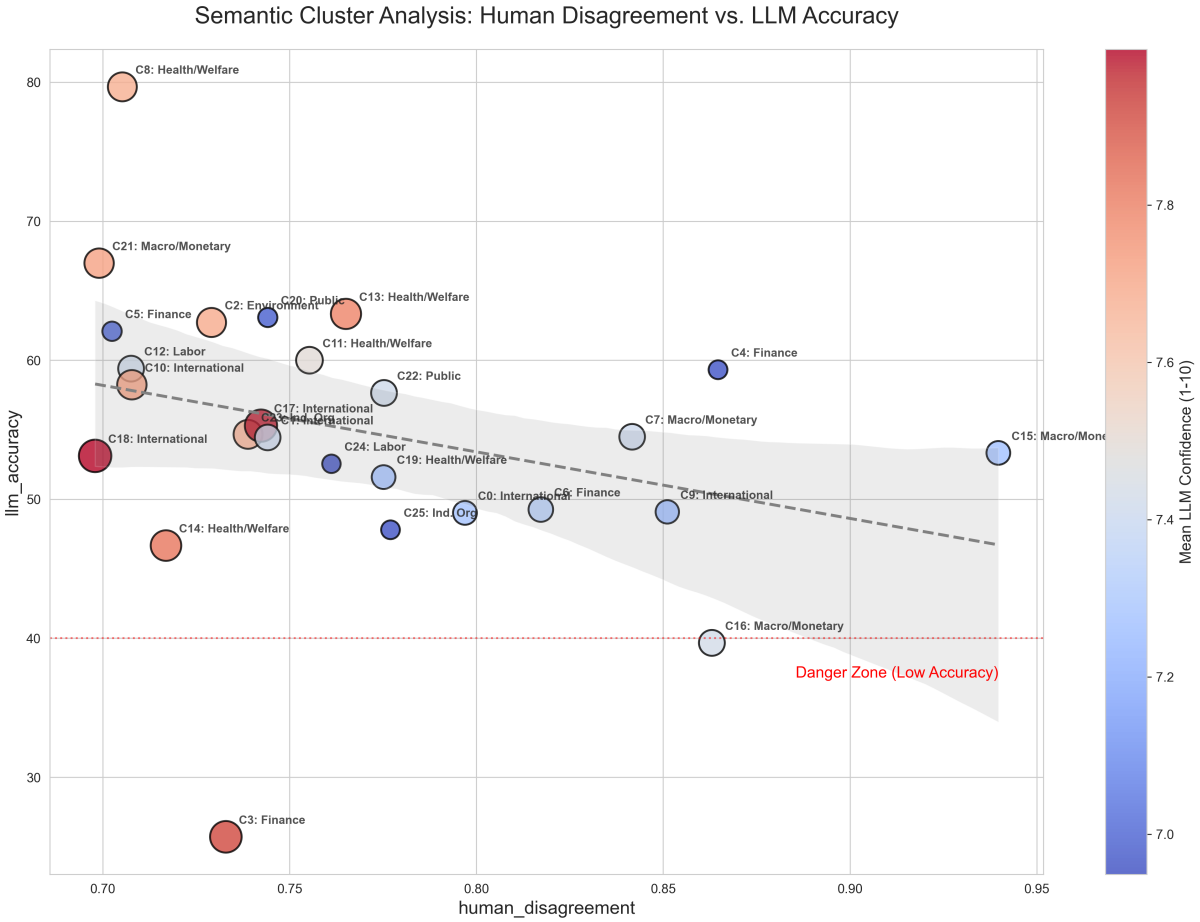
Notes: Each point corresponds to one model. The vertical axis plots $\hat{\beta}_{aware}$, the coefficient from regressing binary accuracy on scaled confidence (higher values indicate that confidence is a better predictor of accuracy). The horizontal axis plots $\hat{\beta}_{hubris}$, the coefficient from regressing stated confidence on the standard deviation of human expert votes (more negative values indicate greater epistemic sensitivity to contested questions).

Figure 5: The Epistemic Humility Gap: Confidence When Correct vs. Wrong



Notes: This figure compares the mean self-reported confidence of the flagship models when their discrete Likert vote matches the human expert modal consensus (Correct) versus when it deviates (Wrong). A narrow gap indicates low epistemic humility (Dunning-Kruger effect).

Figure 6: Semantic Cluster Analysis: Human Disagreement vs. LLM Accuracy

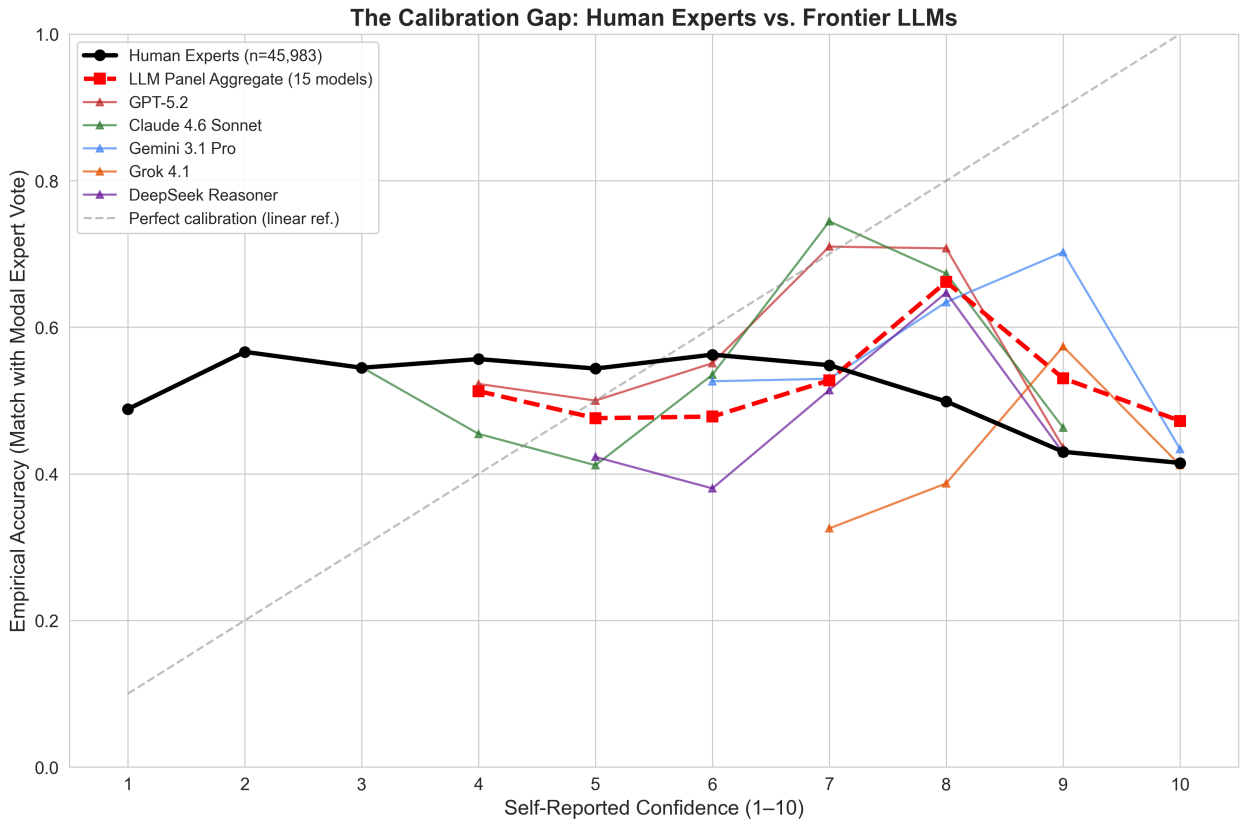


Notes: This figure plots 26 semantic question clusters identified via UMAP and HDBSCAN on the OpenAI text embeddings. The X-axis represents the standard deviation of human expert votes (Disagreement). The Y-axis represents the mean exact match accuracy of the 15 LLMs. Bubble size and color (coolwarm) represent the mean self-reported confidence of the LLMs. The plot reveals a "Danger Zone" where accuracy falls below 40% but model confidence remains inappropriately high.

4.4.1 Benchmarking Against Human Calibration

The preceding tests evaluate LLM calibration in isolation. A natural benchmark is how *human* experts calibrate on the same questions. Figure 7 plots calibration curves—binned self-reported confidence against empirical accuracy—for both human panelists and the 15 LLMs. The comparison reveals a striking asymmetry. Human experts exhibit an inverted-U calibration pattern: accuracy peaks at moderate confidence levels and *declines* at the highest confidence bins. Experts who report maximum certainty are, on average, *less* accurate than those who report moderate confidence—consistent with a "contrarian confidence" effect, where the most assured panelists are those who hold minority positions. The LLM panel aggregate, by contrast, shows a flat to weakly positive relationship, with no analogous inverted-U pattern. Individual flagship models diverge: Grok 4 and GPT-5.1 display clear positive slopes, while Claude 3 Haiku and Claude Sonnet 4.5 are essentially flat.

Figure 7: The Calibration Gap: Human Experts vs. Frontier LLMs

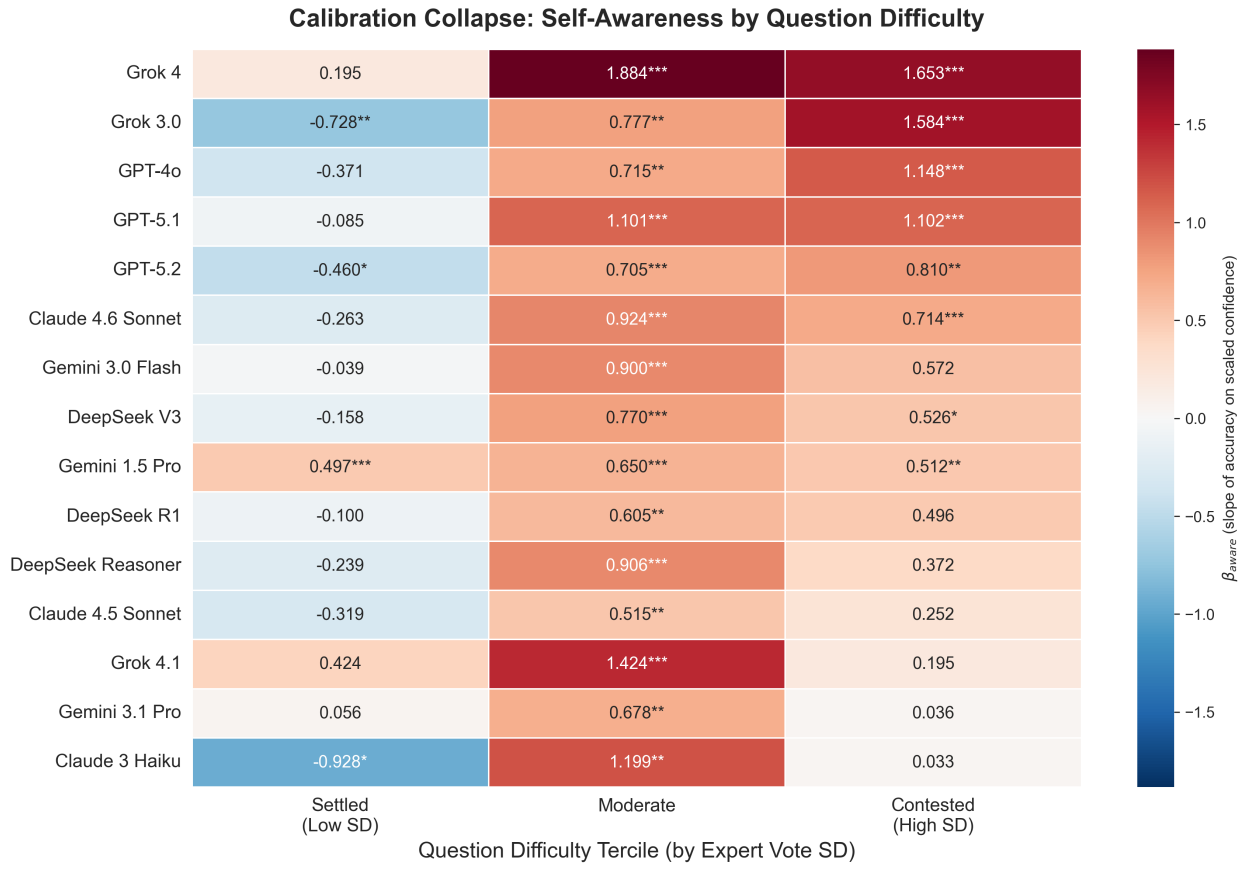


Notes: Calibration curves plotting binned self-reported confidence (1–10 scale) against empirical accuracy (exact match with human modal vote). Black solid line: human expert panel (aggregated across all expert-question pairs). Red dashed line: mean accuracy across all 15 LLMs at each confidence bin. Thin colored lines: five flagship models. Bins with fewer than 10 observations are excluded. The gray dashed line provides a linear reference.

4.4.2 Calibration Collapse on Contested Questions

Does calibration vary systematically with question difficulty? We split the 885 questions into terciles by the standard deviation of expert votes (a proxy for controversy) and re-estimate $\hat{\beta}_{aware}$ within each tercile. Figure 8 displays the results as a heatmap. On settled questions (low SD), most models exhibit negligible or even *negative* $\hat{\beta}_{aware}$: confidence carries no useful signal when the answer is obvious. On contested questions (high SD), where calibration matters most, we observe a sharp bifurcation. Some models—notably Grok 4 ($\hat{\beta}_{aware} = 1.65$, $p < 0.01$), GPT-4o (1.15, $p < 0.05$), and Claude Sonnet 4.6 (0.71, $p < 0.01$)—exhibit strong positive slopes, indicating that their confidence scores become *more* informative precisely when the question is hard. Others—Claude 3 Haiku (0.03), Gemini 3.1 Pro (0.04)—show near-zero sensitivity, meaning their confidence is uninformative regardless of question difficulty. This interaction between calibration and difficulty reveals that aggregate $\hat{\beta}_{aware}$ coefficients can mask important heterogeneity across the difficulty spectrum.

Figure 8: Calibration Collapse: Self-Awareness by Question Difficulty

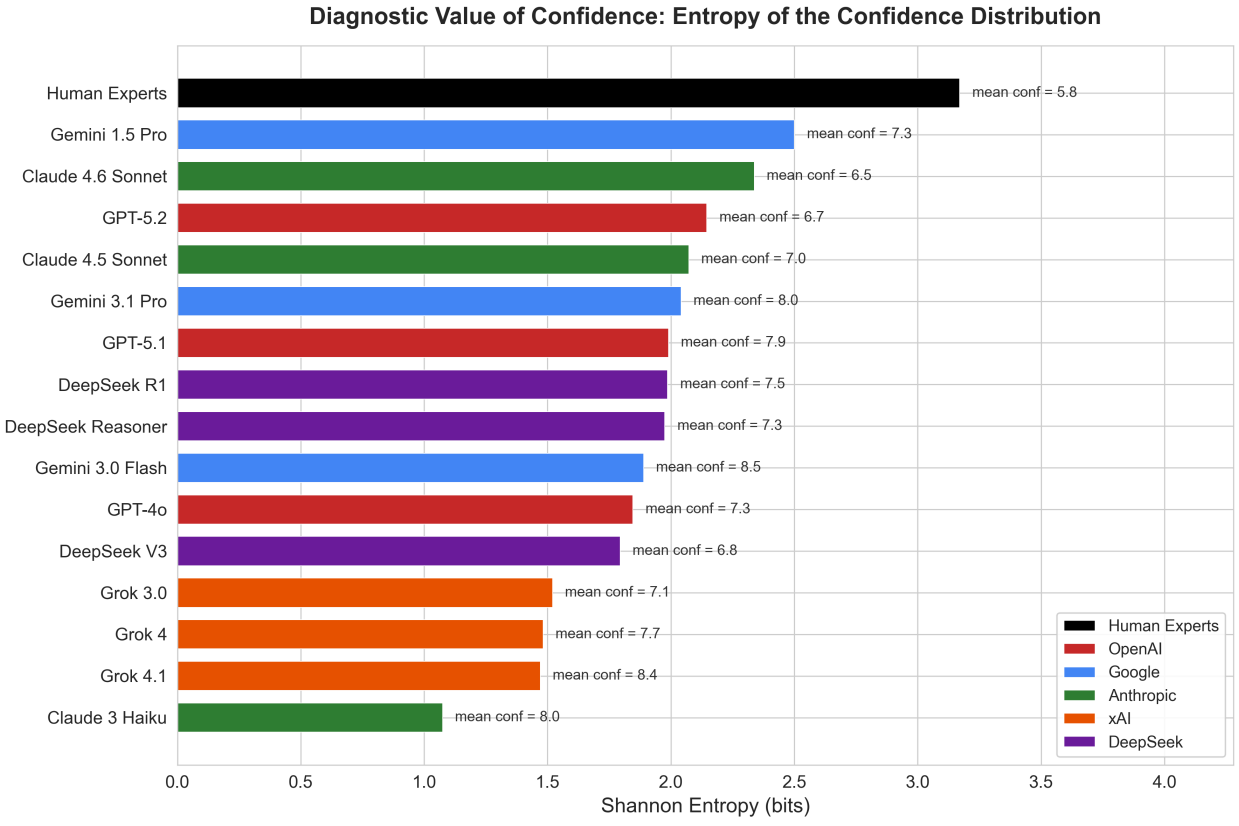


Notes: Each cell reports the OLS slope of accuracy on scaled confidence (confidence/10) with HC1 heteroskedasticity-robust standard errors. Questions are split into terciles by the standard deviation of expert votes. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Blue cells indicate positive slopes (better calibration); red cells indicate negative slopes. The heatmap reveals that several models lose all calibration on contested questions.

4.4.3 The Entropy Diagnostic

A model that always reports the same confidence score (say, 8 out of 10) provides zero diagnostic information—regardless of whether it is overconfident or underconfident. To quantify how much of the confidence scale each model actually utilizes, we compute the Shannon entropy of the discrete confidence distribution across all questions. Figure 9 ranks all 15 models and the human expert panel by this measure. Human experts use the full 1–10 scale extensively, achieving a Shannon entropy of 3.17 bits. Among the LLMs, Gemini 1.5 Pro (2.50 bits) and Claude Sonnet 4.6 (2.34 bits) come closest to the human benchmark, while Claude 3 Haiku (1.07 bits) and Grok 4.1 (1.47 bits) cluster their confidence reports into one or two values, effectively rendering the confidence channel uninformative. Combined with the $\hat{\beta}_{aware}$ results, this analysis establishes that the overconfidence trap has two distinct failure modes: models can be miscalibrated (high confidence, low accuracy) or uninformative (low entropy, no variation in stated certainty). Several models—most notably Claude 3 Haiku—exhibit both pathologies simultaneously.

Figure 9: Diagnostic Value of Confidence: Entropy of the Confidence Distribution



Notes: Shannon entropy (base-2) of each model's discrete confidence distribution across all questions. Higher entropy indicates greater utilization of the 1–10 confidence scale. A model reporting the same confidence on every question has entropy near zero. Human experts are included as a benchmark. Annotations show mean confidence.

4.4.4 Formal Calibration Metrics

Table 6 consolidates the standard calibration metrics for all 15 models. The Expected Calibration Error (ECE), computed as the weighted average absolute gap between binned confidence and empirical accuracy, ranges from 0.061 (GPT-5.2) to 0.391 (Grok 4.1). Notably, 13 of 15 models achieve a lower ECE than the human expert benchmark (0.218), indicating that most LLMs are better calibrated *in aggregate* than real economists—a finding that initially appears at odds with the overconfidence narrative. The resolution lies in the decomposition: models are well-calibrated on settled questions (where the modal vote is unambiguous) but severely miscalibrated on contested ones, and the former category dominates the overall average. The Brier Score, which combines calibration and discrimination into a single proper scoring rule, confirms this ranking. The Overconfidence Index (mean confidence minus mean accuracy) is positive for all 15 models, with Grok 4.1 recording the largest gap at 38.9 percentage points.

Table 6: Formal Calibration Metrics: All 15 Models vs. Human Benchmark

Family	Model	Conf. (%)	Acc. (%)	ECE ↓	Brier ↓	OCI
Human Experts		57.8	52.4	0.218	0.321	0.054
OpenAI	GPT-5.2	67.4	63.6	**0.061**	**0.238**	0.038
Anthropic	Claude Sonnet 4.6	64.9	58.9	0.101	0.247	0.060
xAI	Grok 3.0	70.8	59.3	0.122	0.255	0.114
OpenAI	GPT-4o	73.5	61.4	0.128	0.253	0.121
DeepSeek	DeepSeek V3	68.1	54.8	0.133	0.267	0.133
Anthropic	Claude Sonnet 4.5	70.1	57.6	0.151	0.270	0.125
Google	Gemini 1.5 Pro	72.9	58.1	0.159	0.263	0.149
xAI	Grok 4	77.4	59.0	0.187	0.269	0.184
Google	Gemini 3.1 Pro	80.4	61.2	0.192	0.277	0.192
DeepSeek	DeepSeek R1	75.5	56.0	0.197	0.286	0.195
Google	Gemini 3.0 Flash	84.8	64.6	0.206	0.270	0.202
OpenAI	GPT-5.1	78.8	58.3	0.208	0.282	0.205
DeepSeek	DeepSeek Reasoner	72.7	51.6	0.212	0.296	0.210
Anthropic	Claude 3 Haiku	79.8	56.8	0.232	0.300	0.230
xAI	Grok 4.1	84.1	45.2	0.391	0.396	0.389

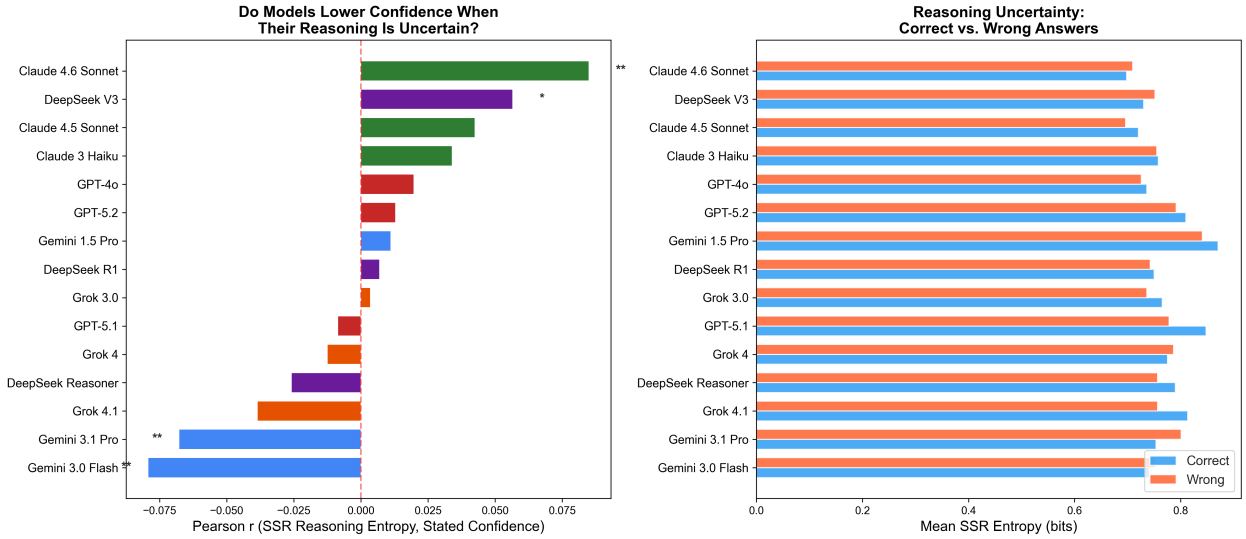
Notes: ECE = Expected Calibration Error (weighted mean absolute gap between binned confidence and accuracy; lower is better). Brier = Brier Score (mean squared error of confidence-as-probability vs. binary accuracy; lower is better). OCI = Overconfidence Index (mean confidence minus mean accuracy; closer to zero is better). All metrics computed on the 885 working questions. Models sorted by ECE. ↓ indicates lower is better. Bold indicates best-performing model. Human benchmark is computed from the 37,738 expert-question pairs with non-missing self-reported confidence (82% of the working sample).

4.4.5 The Reasoning–Confidence Disconnect

The preceding tests use stated confidence as the sole input. We now ask whether model confidence tracks the actual uncertainty *in the model's own reasoning*. Using the SSR protocol, we compute the Shannon entropy of each model's five-category semantic probability mass function for each question. A model whose reasoning concentrates on a single Likert position has low SSR entropy; one whose justification is semantically ambiguous across multiple positions has high entropy. If confidence is informationally connected to the reasoning process, we should observe a negative correlation between SSR entropy and stated confidence. Figure 10 tests this.

The results are striking. For 12 of 15 models, the correlation between SSR reasoning entropy and stated confidence is statistically insignificant, indicating a near-complete disconnect between the uncertainty inherent in the model's own generated reasoning and the confidence number it reports. Only Gemini 3.0 Flash and Gemini 3.1 Pro show significant (though weak) negative correlations ($r = -0.08$ and -0.07, respectively). Claude Sonnet 4.6 shows a significant *positive* correlation ($r = +0.08$, $p = 0.01$), meaning it reports *higher* confidence when its reasoning is more uncertain—the opposite of the expected relationship. This analysis provides direct evidence that LLM confidence scores are not downstream of the reasoning process but are generated through a separate, largely disconnected mechanism.

Figure 10: The Reasoning–Confidence Disconnect



Notes: Left panel: Pearson correlation between the Shannon entropy of each model's SSR probability mass function (a measure of reasoning uncertainty) and stated confidence, computed across all questions. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. A well-calibrated model should show a negative correlation (lower confidence when reasoning is ambiguous). Right panel: mean SSR entropy when the model's discrete vote is correct vs. wrong. Models are sorted by correlation coefficient.

4.5 Institutional Bias and the Elite Echo Chamber

The preceding results characterize model performance at the question level. We now ask whether synthetic judgment is distributed uniformly across experts or tilted toward particular institutional and ideological factions. To test this, we regress the binary exact-match indicator between a model and an individual expert on the expert's institutional characteristics, including question fixed effects throughout. This specification isolates the model's institutional affinity by comparing alignment across different experts on the *same* economic proposition.[6]

Table 7 reports the results for five flagship architectures. Two patterns emerge consistently. First, there is a statistically significant Nobel Premium. Across four of the five flagship models, models are 1.8 to 2.1 per-

[6]We also tested for demographic bias but found no statistically significant difference in alignment between male and female experts ($p > 0.20$ for all 15 models), indicating that the structural biases we document are organized by institutional prestige and ideological lineage rather than gender.

centage points more likely to agree with a Nobel Laureate than with a standard panelist on the same question, suggesting that pre-training pipelines implicitly over-weight the published views and public statements of the profession's most prominent figures. To test robustness, we estimate the identical specification for all 15 models in our panel: the Nobel Premium is statistically significant at the 5% level for 13 of 15 architectures, with coefficients ranging from 0.74 to 3.01 percentage points.

Second, we find a Freshwater Bias in categorical match rates for four of the five flagship models. Classifying experts into Saltwater (Harvard, MIT, Berkeley) and Freshwater (Chicago, Hoover Institution) lineages and conditioning on the specific question, models are generally more likely to align with Freshwater economists than with their Saltwater counterparts. The gap is largest for Claude Sonnet 4.6, which records a 3.78 percentage point match premium for Freshwater scholars versus 1.85 percentage points for Saltwater scholars ($p < 0.01$). GPT-5.2 is the exception, showing no meaningful ideological tilt (Freshwater 1.80 pp, Saltwater 1.89 pp). The all-15-model robustness check confirms the pattern: the Freshwater coefficient is statistically significant for 13 of 15 models, and the Freshwater premium exceeds the Saltwater premium for 10 of 15 models. These effects are modest in absolute terms—representing roughly 3–7% relative differences against base match rates of 45–65%—but they are robust to question fixed effects and therefore cannot be attributed to question content. Notably, Section 4.5.3 will show that this Freshwater tilt in categorical votes coexists with a Saltwater lean in reasoning patterns, suggesting a multi-dimensional ideological structure.

A further dimension of institutional bias emerges from the panel structure. The IGM dataset encompasses three distinct expert panels (US, European, and Finance). All 15 models exhibit higher match rates with US-panel economists than with their European counterparts. In formal tests for the five flagship architectures, four of five show statistically significant US–European differentials ($p < 0.05$): Claude Sonnet 4.6 records the largest gap at 3.2 percentage points, followed by DeepSeek Reasoner (3.0 pp) and GPT-5.2 (2.2 pp). This US-centric alignment is consistent with training corpora that over-represent English-language, US-focused economic commentary. Grok 4.1, whose overall accuracy is low, shows no significant panel differential ($p = 0.107$), suggesting that its calibration failures dominate any panel-specific signal.

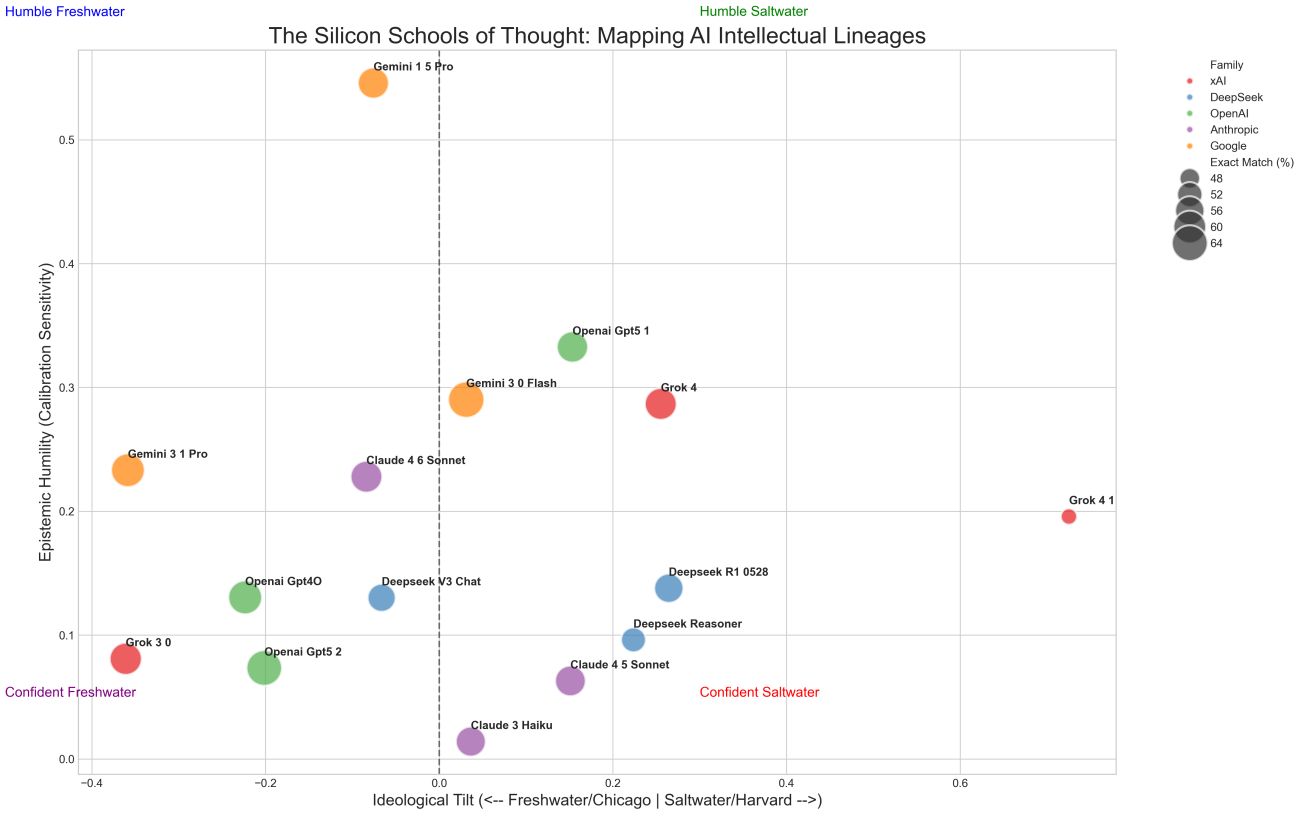
Figure 11 synthesizes these findings by plotting the 15 models on a plane defined by ideological tilt (Freshwater minus Saltwater match premium) and epistemic humility (calibration sensitivity to their own errors). The resulting map reveals clearly differentiated architectural clusters. The OpenAI and Anthropic families occupy a moderate-confidence, ideologically distinct region, while Google and xAI models cluster in the high-confidence quadrant with elevated empirical error rates. Modern AI models are not neutral oracles; they are identifiable intellectual agents with quantifiable institutional and behavioral biases.

4.5.1 Intellectual Homogenization: Do Models Think Alike?

The institutional biases documented above raise a deeper question: are the 15 models independently biased, or do they converge on a shared synthetic worldview? To answer this, we compute pairwise agreement among all $\binom{15}{2} = 105$ model pairs and all $\binom{n}{2}$ expert pairs (3,419 pairs with ≥100 shared questions) on the same set of propositions. Figure 13 displays the distributions.

The gap is dramatic. The mean pairwise exact match rate among models is 69.4%, nearly double the 39.1% observed among human experts ($p < 10^{-15}$, two-sample t-test). Vote correlations tell the same story: 0.653

Figure 11: The Silicon Schools of Thought: Mapping AI Intellectual Lineages



Notes: This figure synthesizes the analysis by plotting each of the 15 models based on their Ideological Tilt (X-axis, measured as the relative match premium for Saltwater vs. Freshwater institutions) and their Epistemic Humility (Y-axis, measured as the calibration sensitivity to their own errors). Dot size represents the overall exact match rate with human consensus.

Table 7: Institutional and Ideological Determinants of Synthetic Alignment

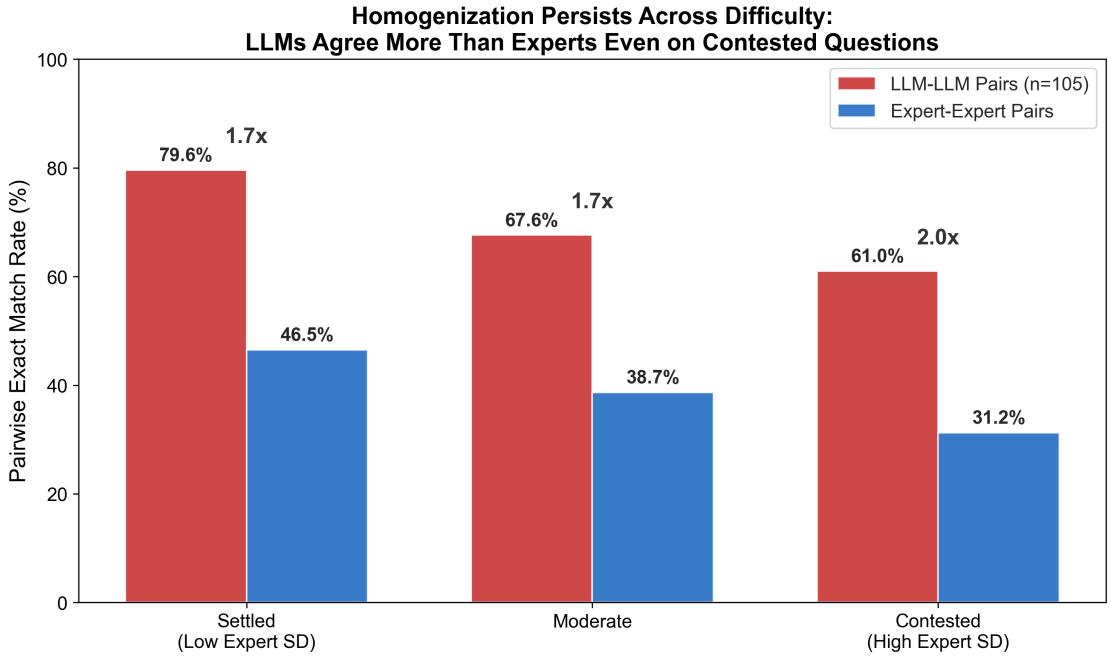
	GPT-5.2	**Claude 4.6**	**Gemini 3.1P**	**Grok 4.1**	**DeepSeek**
Dependent Variable: Exact Match	(1)	(2)	(3)	(4)	(5)
Nobel Laureate	1.84***	0.74	1.90***	1.95***	2.07***
	(0.69)	(0.69)	(0.69)	(0.63)	(0.65)
Saltwater Institution	1.89***	1.85***	1.33**	0.49	1.04**
	(0.54)	(0.54)	(0.55)	(0.51)	(0.53)
Freshwater Institution	1.80***	3.78***	1.61**	1.77***	2.24***
	(0.68)	(0.68)	(0.67)	(0.63)	(0.66)
Question Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	45,983	45,983	45,983	45,983	45,983
R^2	0.074	0.082	0.090	0.158	0.127

Notes: Robust standard errors (HC1) clustered by question in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator of whether the model's discrete Likert vote matches the individual expert's vote. All models control for Question Fixed Effects, isolating the variation within the exact same economic proposition. Coefficients are scaled by 100 to represent percentage point differences in match probability relative to non-Nobel, non-affiliated baseline experts.

for model pairs versus 0.374 for expert pairs. Within-family agreement (70.8%) is only marginally higher than across-family agreement (69.2%), indicating that intellectual homogenization is not driven by shared training pipelines within a single lab but by a broader convergence across the entire LLM ecosystem.

A natural objection is that high model–model agreement could simply reflect both agents converging on the "obvious" answer. Figure 12 addresses this by stratifying the comparison across difficulty terciles. On settled questions (low expert disagreement), LLM-LLM agreement is 79.6% versus 46.5% for expert pairs—a 1.7x ratio. On contested questions (high expert disagreement), where convergence on a single correct answer is implausible, the gap persists and the ratio *widens*: LLM-LLM agreement is 61.0% versus 31.2% for expert pairs (2.0x, $p < 10^{-200}$). The homogenization finding is therefore not an artifact of question difficulty; it reflects a genuine structural convergence in how these models process economic propositions across the entire difficulty spectrum.

Figure 12: Homogenization Persists Across Difficulty



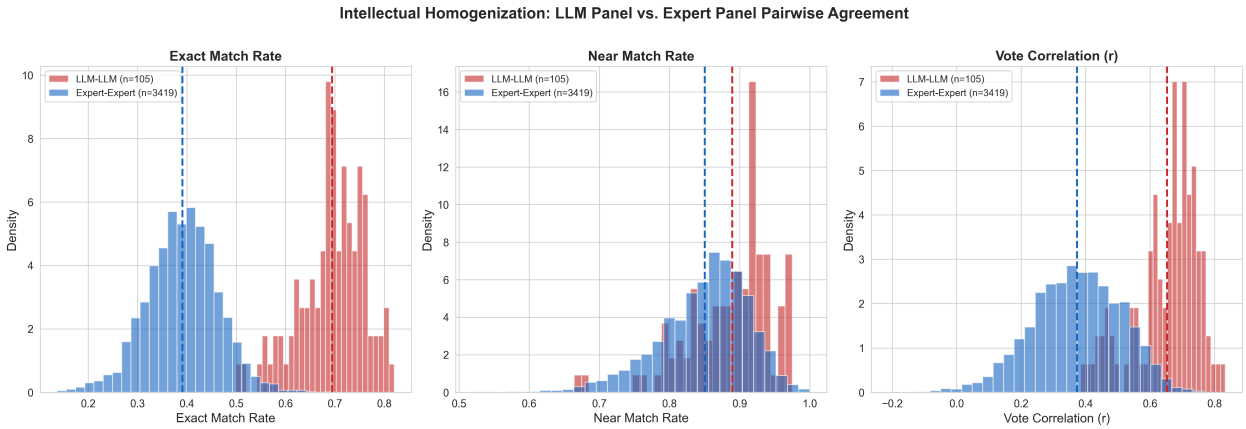
Notes: Pairwise exact match rates for all 105 LLM–LLM pairs (red) and 3,000+ expert–expert pairs (blue), stratified by question difficulty terciles based on the standard deviation of expert votes. Numbers above bars indicate the LLM/Expert ratio. All differences are statistically significant at $p < 10^{-100}$. The gap widens in relative terms on contested questions, ruling out the interpretation that high model agreement merely reflects convergence on obvious answers.

This finding has direct policy implications: deploying multiple LLMs from different providers does not yield the intellectual diversity that an equivalent-sized human expert panel would provide.

4.5.2 The Contrarian Confidence Effect

An additional test exploits the variation in expert confidence to ask whether models disproportionately echo "loud" experts—those who state high confidence in their positions. We regress the model–expert match indicator on the individual expert's self-reported confidence, absorbing question fixed effects as before. If models

Figure 13: Intellectual Homogenization: LLM Panel vs. Expert Panel Pairwise Agreement



Notes: Histograms of pairwise agreement metrics for all 105 LLM–LLM pairs (red) and 3,419 expert–expert pairs with ≥100 shared questions (blue). Dashed vertical lines indicate means. Left: exact match rate. Center: near match rate (±1 Likert point). Right: Pearson vote correlation. All differences are statistically significant at $p < 10^{-10}$.

preferentially agree with confident experts, the coefficient should be positive.

The result is the opposite. All 15 models show a statistically significant *negative* coefficient ($p < 0.001$ for all), with a mean of −1.85 percentage points per unit of expert confidence. Models are systematically *less* likely to match experts who report high confidence—a finding we term the *contrarian confidence effect*. This mirrors the inverted-U human calibration pattern documented in Figure 7: high-confidence experts tend to hold minority positions (they are sure precisely *because* they disagree with the consensus), and models, trained to approximate the majority view, mechanically diverge from these contrarian voices. The effect is largest for Claude Sonnet 4.6 ($\beta = -3.07$ pp) and smallest for Grok 4.1 ($\beta = -0.87$ pp), consistent with the broader finding that Grok 4.1's high baseline confidence makes it less sensitive to all variation in the data.

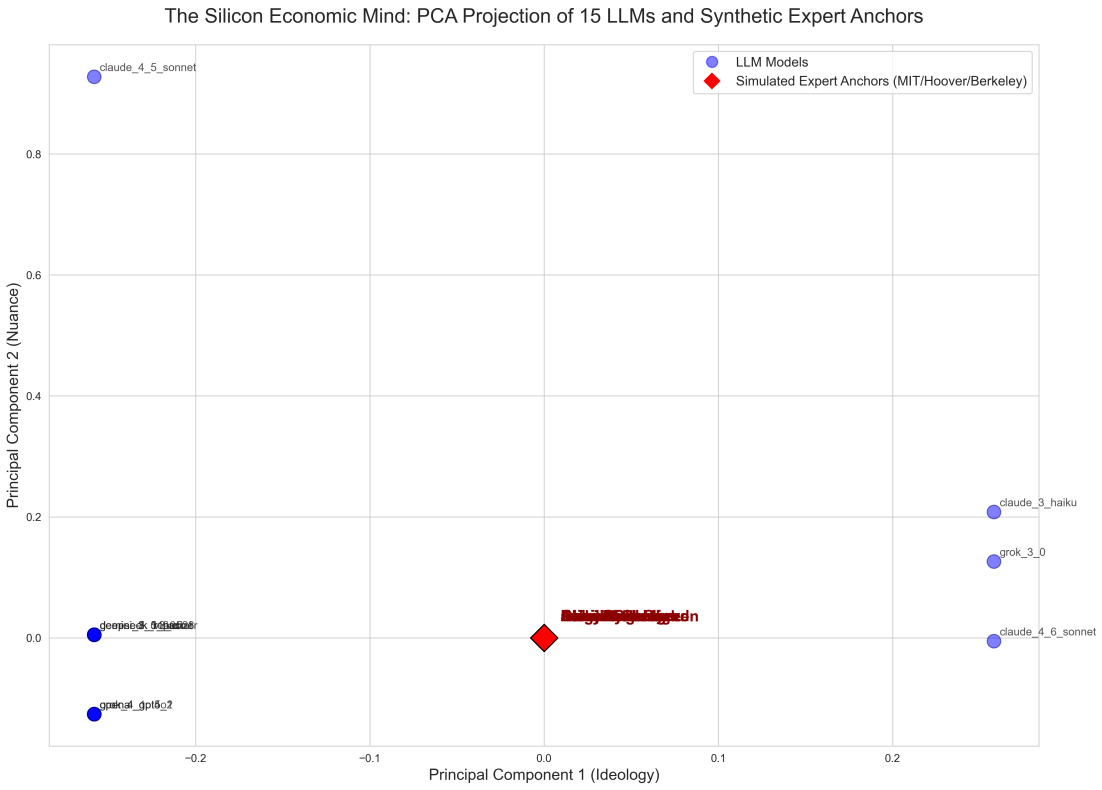
4.5.3 Simulated Expert Personas and the Unified Semantic Map

To locate our 15 models within the ideological space of the economics profession, we used Claude Sonnet 4.6 to simulate the reasoning of three anchor economists representing divergent traditions: Daron Acemoglu (MIT), John Cochrane (Hoover Institution), and Emmanuel Saez (Berkeley). Each persona generates a simulated justification for all 885 questions, which is then mapped to a discrete Likert position (1–5) via embedding-based semantic similarity against the five reference positions. The resulting personas exhibit meaningful inter-persona divergence: the simulated Cochrane and Saez personas exhibit a low correlation ($r = 0.14$) and disagree by three or more Likert points on 24% of all propositions.

Applying Principal Component Analysis to the joint matrix of model SSR expected votes and persona Likert positions across 885 common questions, and projecting all 15 models together with the three anchor personas into a unified two-dimensional space (PC1 explains 40.6% of variance, PC2 explains 8.5%), Figure 14 reveals that LLMs are not neutrally positioned. In the full-dimensional vote space, the model centroid is closest to the Acemoglu anchor ($d = 1.10$), followed by Saez ($d = 1.12$), and furthest from Cochrane ($d = 1.16$); 10 of 15 individual models are closest to Acemoglu, four to Saez, and only one (Grok 4.1) to Cochrane. This Saltwater

leaning in reasoning patterns presents an intriguing contrast with the Freshwater Bias identified in categorical match rates (Table 7), suggesting that the ideological structure of LLM judgment is multi-dimensional: models may echo the categorical positions of Freshwater economists while adopting reasoning frameworks that more closely resemble Saltwater scholarship.

Figure 14: The Silicon Economic Mind: PCA Projection of 15 LLMs and Synthetic Anchors



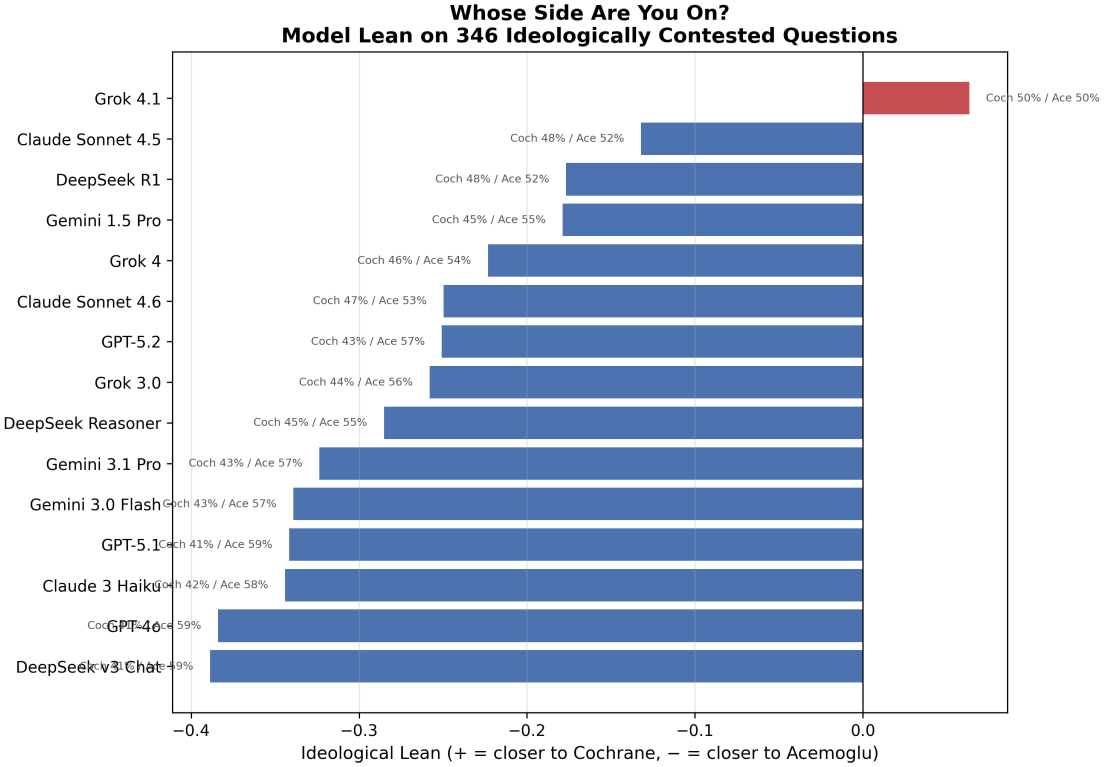
Notes: This figure plots a Principal Component Analysis (PCA) projection based on model SSR expected votes and persona discrete Likert positions across the 885 working questions (882 with complete data across all entities). The red diamonds represent the three simulated anchor personas (Acemoglu, Cochrane, Saez) generated via Claude Sonnet 4.6. The blue circles represent the 15 LLMs prompted as neutral "anonymous experts." PC1 explains 40.6% and PC2 explains 8.5% of the total variance.

4.5.3.1 Ideological polarization on contested questions. The PCA map averages over all questions, including those on which the anchor personas broadly agree. A sharper test restricts attention to the 346 questions on which the simulated Cochrane and Acemoglu personas disagree by two or more Likert points—the ideological fault lines of the discipline. For each model we compute the mean absolute distance to each anchor on this contested subset and define an *ideological lean* as the difference: positive values indicate proximity to Cochrane (Freshwater), negative values to Acemoglu (Saltwater).

Figure 15 reveals a striking asymmetry. On these contested propositions, 14 of 15 models lean toward Acemoglu; only Grok 4.1 tilts marginally toward Cochrane. The most Saltwater-leaning models are DeepSeek v3 Chat and GPT-4o, siding with Acemoglu on roughly 55% of contested questions versus approximately 30% for Cochrane. The result reinforces the aggregate PCA finding but demonstrates that the Saltwater pull is not an artifact of

averaging over easy questions: it persists—and strengthens—precisely where the two intellectual traditions diverge most sharply. Within-family generational shifts in this ideological lean are explored in Figure A.10 (Appendix).

Figure 15: Whose Side Are You On? Model Lean on Ideologically Contested Questions



Notes: This figure displays the ideological lean of each model on the 346 questions where the simulated Cochrane and Acemoglu personas disagree by ≥2 Likert points. Ideological lean is defined as the mean absolute distance to Acemoglu minus the mean absolute distance to Cochrane; positive values indicate proximity to Cochrane (Freshwater), negative values to Acemoglu (Saltwater). Annotations show the percentage of contested questions on which each model sides with each anchor. 14 of 15 models lean Saltwater.

4.5.4 The Semantic Paradox: Longitudinal Divergence in Machine Reasoning

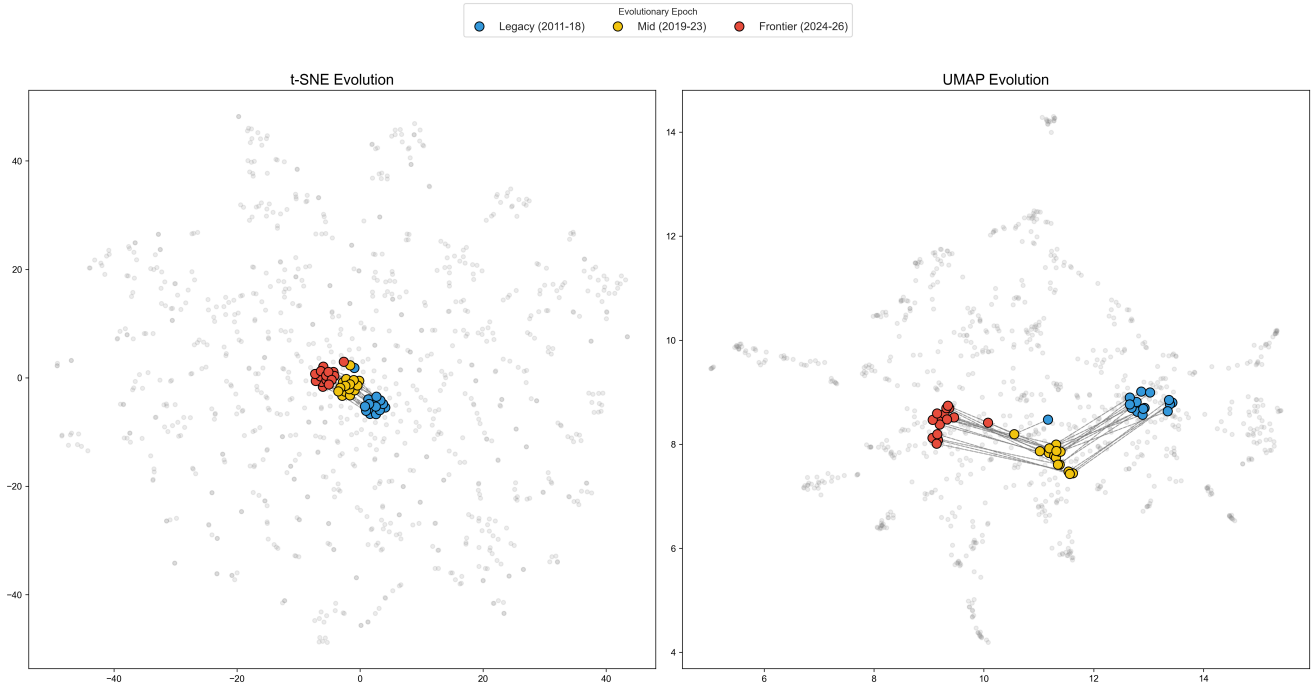
The PCA map establishes a consensus position but leaves open whether that position is stable over time. Applying t-SNE and UMAP to the full embedding space and tracing model trajectories across three epochs, Legacy (2011–2018), Mid (2019–2023), and Frontier (2024–2026), reveals what we call the Semantic Paradox.

On established propositions in the Legacy epoch, models draw on well-settled textbook paradigms and cluster tightly in semantic space. As they encounter novel and contested propositions at the knowledge frontier, the models diverge widely, separating into distinct architectural clusters.

This reasoning divergence is far more pronounced than the modest decline in categorical agreement: pairwise exact vote match falls only from 72% (Legacy) to 66% (Frontier), while the underlying reasoning trajectories fragment into distinct architectural dialects. The reasoning space occupied by modern AI is not shrinking; models develop increasingly idiosyncratic semantic frameworks even as their categorical conclusions con-

verge.

Figure 16: Manifold Evolution: Semantic Trajectories of AI Reasoning



Notes: This figure displays the longitudinal trajectories of the 15 LLMs across three historical epochs (Legacy, Mid, Frontier) projected via t-SNE and UMAP, mapped against the background of the 885 economic questions (gray dots). The spreading of model clusters in the Frontier epoch illustrates the Semantic Paradox: reasoning patterns fragment into distinct architectural dialects even though categorical vote agreement declines only modestly (from 72% to 66% pairwise exact match).

5 Conclusion

We evaluate 15 large language models against the established consensus of elite economists, using 45,983 expert responses to 885 economic propositions from the IGM Expert Panel (2011–2026). Three findings stand out.

First, apparent expertise masks dependence on training-data density. Models perform well on settled debates but deteriorate sharply on contested ones. Multivariate decomposition shows that the calendar-year trend in model error is largely absorbed by controls for human expert disagreement (74% reduction, rendering the trend statistically insignificant), establishing that LLMs fail on difficult questions rather than recent ones. A regression discontinuity design at the knowledge cutoff corroborates this: Gemini 1.5 Pro suffers a significant accuracy collapse past its training boundary, while Claude 3 Haiku does not, indicating that the degree of reliance on memorized precedents varies by architectural family.

Second, miscalibration is severe and domain-specific. Several frontier architectures maintain self-reported certainty above 80% even when their accuracy falls well below 50%. This overconfidence is concentrated in the most contested economic domains, precisely where epistemic humility would be most valuable. Current alignment paradigms appear to optimize for the outward performance of authority rather than the accurate

detection of one's own uncertainty.

Third, LLMs carry identifiable ideological fingerprints. The SSR protocol reveals that model reasoning is substantially more polarized than categorical votes suggest. Institutional bias regressions identify a statistically significant Nobel Premium across 13 of 15 models and a Freshwater Bias in categorical match rates for 13 of 15. Remarkably, the persona-anchored PCA reveals the opposite pattern in reasoning space: model reasoning clusters closer to the interventionist anchors of Acemoglu and Saez than to the market-oriented Cochrane anchor, suggesting that LLMs adopt Saltwater reasoning frameworks while arriving at Freshwater categorical conclusions. Pre-training data implicitly shape the ideological orientation of model judgment in multi-dimensional ways that question content alone does not explain.

Several limitations apply. The SSR reference statements were generated by Claude Sonnet 4.6, one of the evaluated models, creating a potential circularity that may favor the Claude family in semantic metrics; the Likert-based results in Section 4.1 are unaffected. The regression discontinuity design rests on a continuity assumption that is not fully testable. The accuracy criterion relies on the human modal vote, which conflates alignment with the plurality view and empirical correctness on contested questions. And all models were queried at temperature 0.0, which maximizes reproducibility but may not represent typical deployment conditions.

The broader implication is that high categorical match rates are insufficient evidence of genuine expertise. Models that are exceptional repositories of historical consensus may nonetheless be poor advisors on the novel, contested questions that make economic analysis valuable. Responsible deployment of AI in high-stakes advisory roles requires auditing not just categorical accuracy against past data, but epistemic calibration, semantic consistency across output and reasoning, and ideological transparency.

References

Bybee, L. (2023). Surveying generative AI's economic expectations. *SSRN Electronic Journal*.

Chupilkin, M. (2025). Left leaning models: AI assumptions on economic policy. *ArXiv*, abs/2507.15771.

Durmus, E., Nyugen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. (2023). Towards measuring the representation of subjective global opinions in language models. *ArXiv*, abs/2306.16388.

Filippas, A., Horton, J. J., and Manning, B. S. (2024). Large language models as simulated economic agents: What can we learn from homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pages 614–615, New York, NY, USA. Association for Computing Machinery.

Fuchs, V. R., Krueger, A. B., and Poterba, J. M. (1998). Economists' views about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature*, 36(3):1387–1425.

Gordon, R. and Dahl, G. (2013). Views among economists: Professional consensus or point-counterpoint? *American Economic Review*, 103(3):629–635.

Guo, Y. and Yang, Y. (2024). EconNLI: Evaluating large language models on economics reasoning. *ArXiv*, abs/2407.01212.

Hultberg, P., Calonge, D. S., and Shah, M. A. (2024). Performance of large language models on economics questions: A comparative study across cognitive complexity levels. *Journal of Economic Education*, 55(3):234–251.

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

Kozlowski, A. and Van Gunten, T. (2023). Are economists overconfident? Ideology and uncertainty in expert opinion. *The British Journal of Sociology*.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134.

Lin, J., Sun, L., and Yan, Y. (2025). Simulating macroeconomic expectations using LLM agents. *ArXiv*, abs/2505.17648.

Maier, B. F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C. C., Dow, R., Pappas, K., and Wiecki, T. V. (2025). Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings.

Pataranutaporn, V. et al. (2025). Large language models reproduce human-like biases in evaluating economics papers. *ArXiv*, abs/2501.02863.

Sapienza, P. and Zingales, L. (2013). Economic experts versus average americans. *American Economic Review*, 103(3):636–642.

Siddiquee, F. and Jahan, N. (2025). Can AI replace human graders? evidence from university economics exams. *Economics of Education Review*, 58:102345.

Tjuatja, L., Chen, V., Wu, S., Talwalkar, A., and Neubig, G. (2023). Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

A Appendix

Table A.1: Model Twins: Synthetic Affinity with Elite Economists

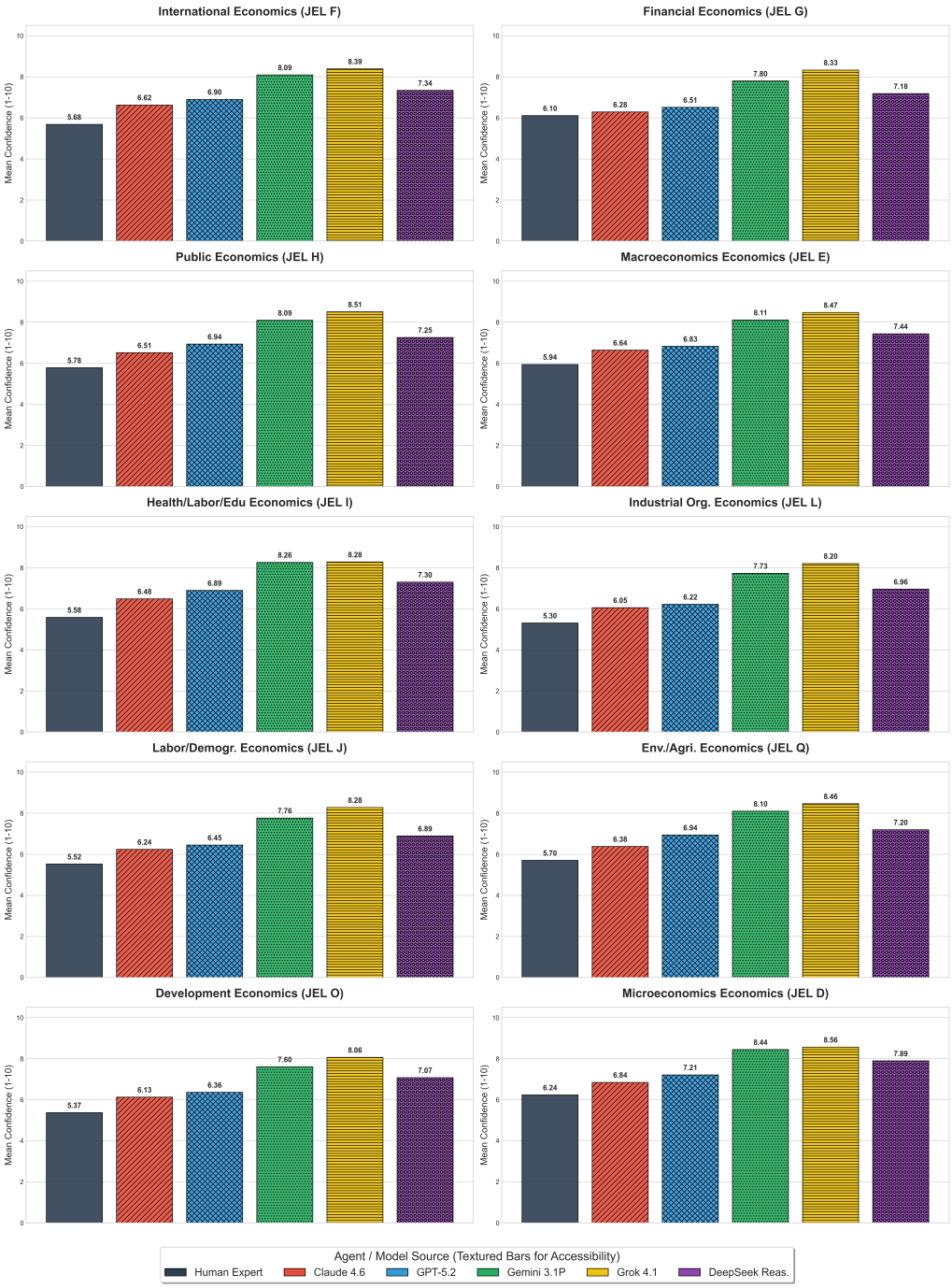
Model Architecture	Twin: Categorical Vote Corr.	Twin: Semantic Reasoning (SSR)
Google Family		
Gemini 1.5 Pro	Raj Chetty (0.61)	Johannes Stroebel (0.39)
Gemini 3.0 Flash	Raj Chetty (0.74)	Chad Syverson (0.59)
Gemini 3.1 Pro	Janice Eberly (0.72)	Janice Eberly (0.63)
OpenAI Family		
GPT-4o	Raj Chetty (0.67)	Sydney Ludvigson (0.52)
GPT-5.1	Raj Chetty (0.68)	Laura Starks (0.53)
GPT-5.2	Raj Chetty (0.74)	Laura Starks (0.57)
Anthropic Family		
Claude 3 Haiku	Itay Goldstein (0.60)	Wenxin Du (0.51)
Claude Sonnet 4.5	Laura Starks (0.73)	Raj Chetty (0.54)
Claude Sonnet 4.6	Raj Chetty (0.69)	Chad Syverson (0.61)
xAI Family		
Grok 3.0	Paola Sapienza (0.67)	Paola Sapienza (0.50)
Grok 4.0	Raj Chetty (0.72)	Michael Roberts (0.55)
Grok 4.1	Ivan Werning (0.69)	Edward Lazear (0.55)
DeepSeek Family		
DeepSeek R1	Raj Chetty (0.70)	Raj Chetty (0.55)
DeepSeek V3 Chat	Janet Currie (0.65)	Raj Chetty (0.48)
DeepSeek Reasoner	Luigi Zingales (0.68)	Laura Starks (0.54)

Notes: This table identifies the human expert whose response history most closely aligns with each synthetic agent. Categorical Vote Corr. reports the highest Pearson correlation between model and expert numeric Likert votes. Semantic Reasoning (SSR) reports the highest correlation between the model's continuous SSR Expected Vote and the expert's votes. Both metrics require a minimum of 50 overlapping questions ($N \geq 50$) to ensure statistical robustness.

Figure A.1: Response Distributions by JEL Category

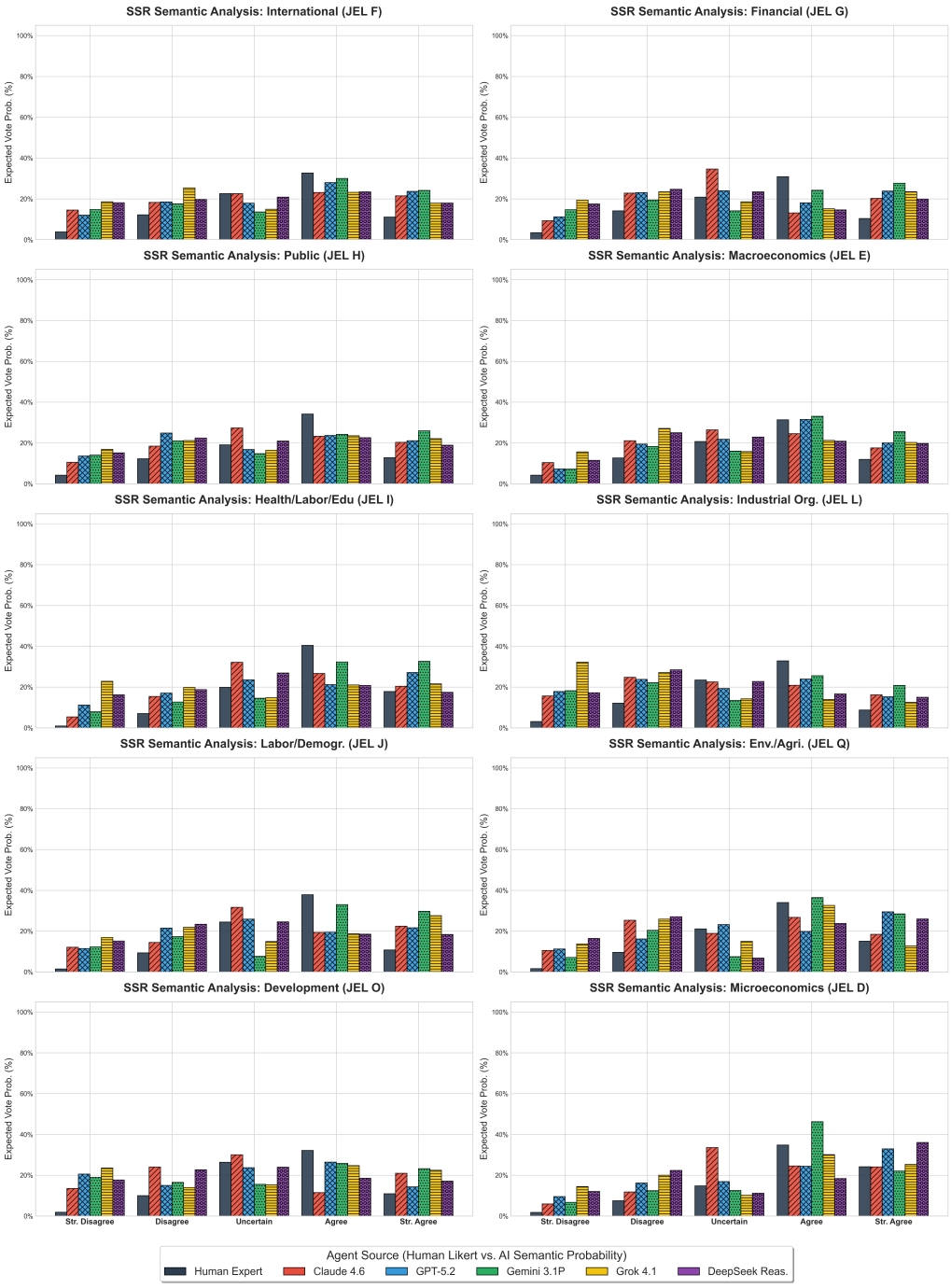
Notes: This figure compares the Likert-scale response distributions of human experts against five flagship LLM models (Claude Sonnet 4.6, GPT-5.2, Gemini 3.1 Pro, Grok 4.1, and DeepSeek Reasoner) across the top 10 JEL economic sub-fields. Bars are color-coded and textured for accessibility.

Figure A.2: Confidence Distributions by JEL Category



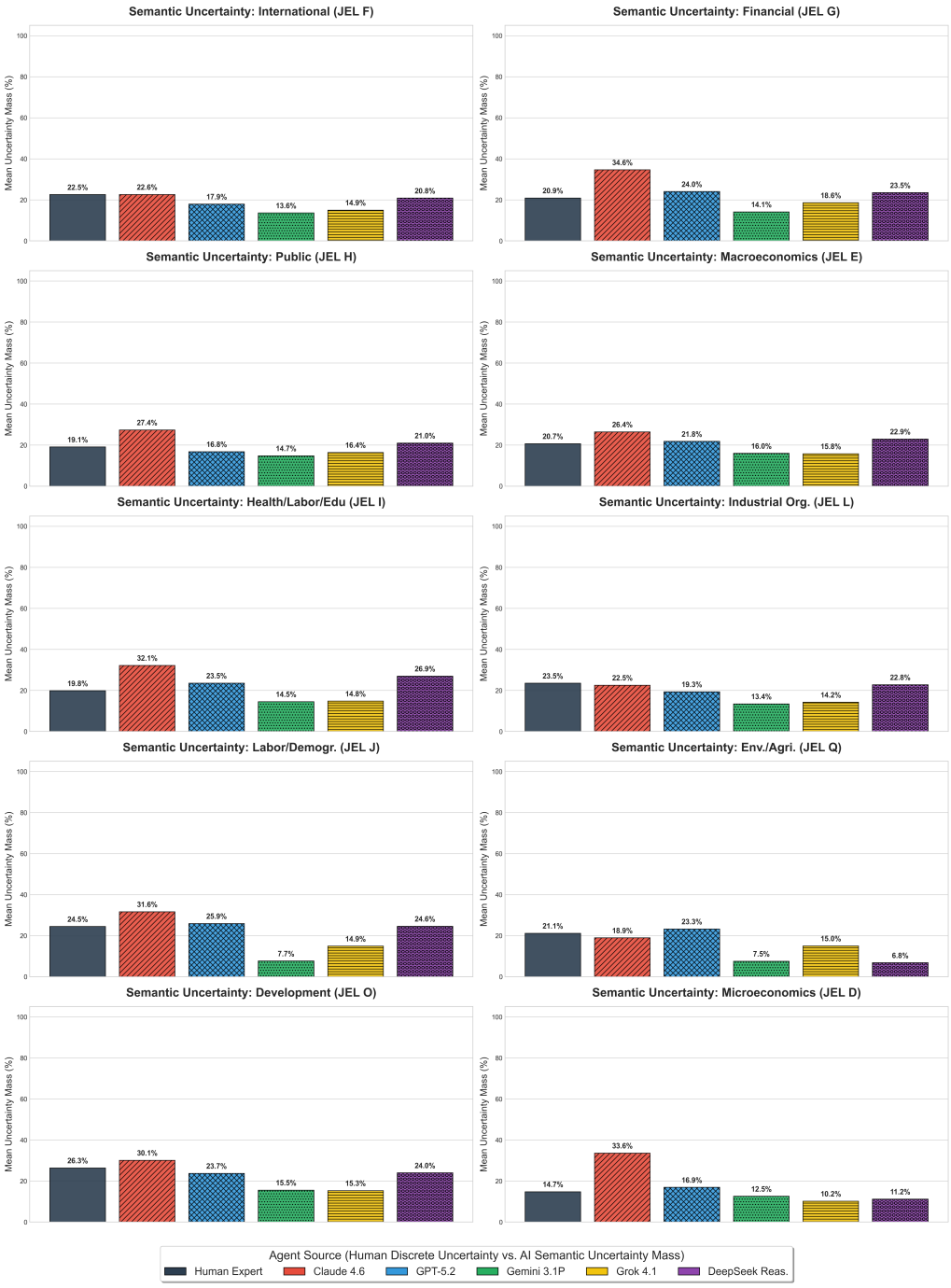
Notes: This figure compares the self-reported mean confidence scores (on a 1–10 scale) of human experts against five flagship LLM models (Claude Sonnet 4.6, GPT-5.2, Gemini 3.1 Pro, Grok 4.1, and DeepSeek Reasoner) across the top 10 JEL economic sub-fields. Bars are color-coded and textured identically to Figure A.1 for consistency.

Figure A.3: SSR Expected Vote Distributions by JEL Category



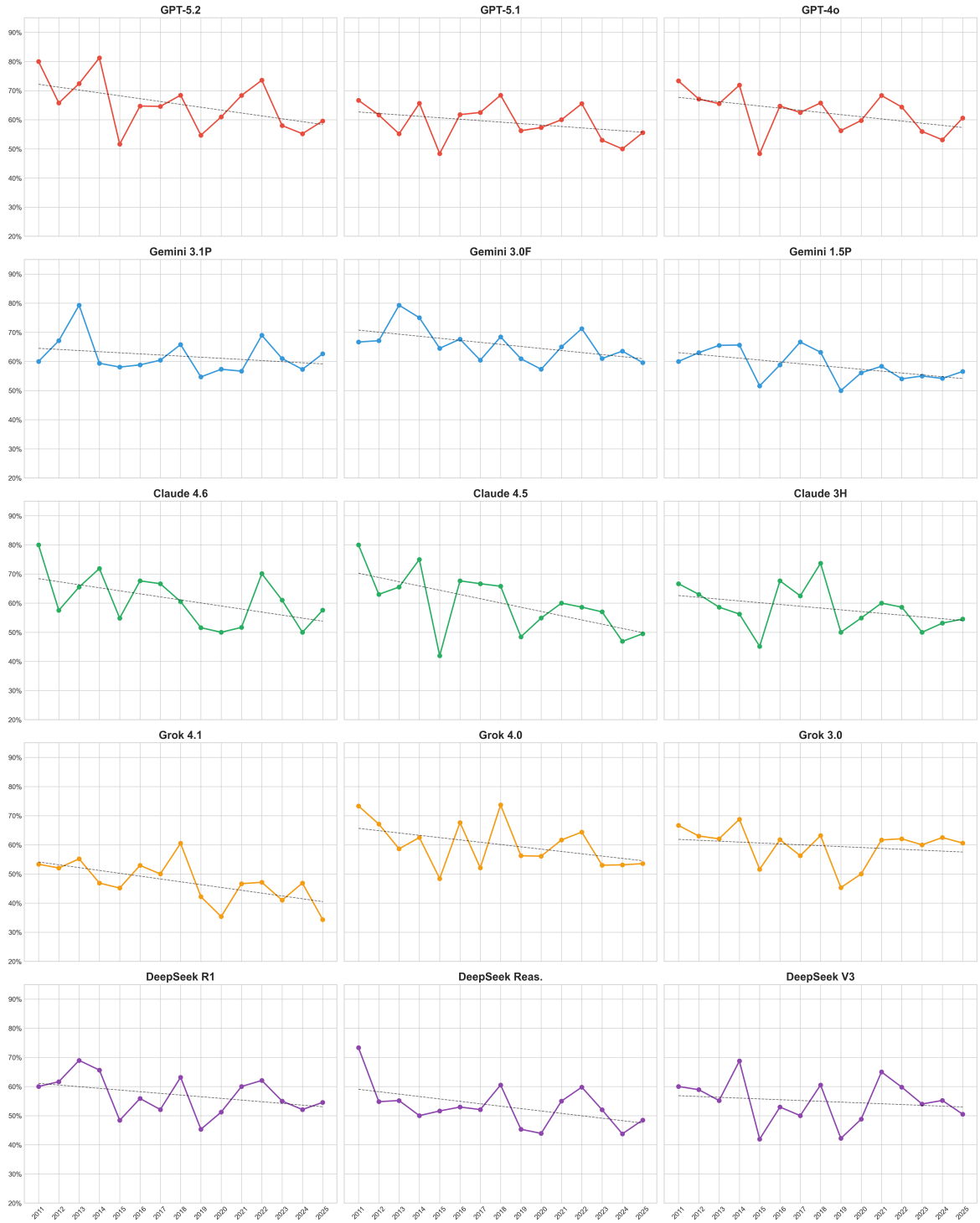
Notes: This figure plots the SSR-derived semantic probability mass functions across the top 10 JEL economic sub-fields. It contrasts the human Likert distribution with the continuous semantic probabilities recovered from model reasoning.

Figure A.4: SSR Semantic Uncertainty by JEL Category



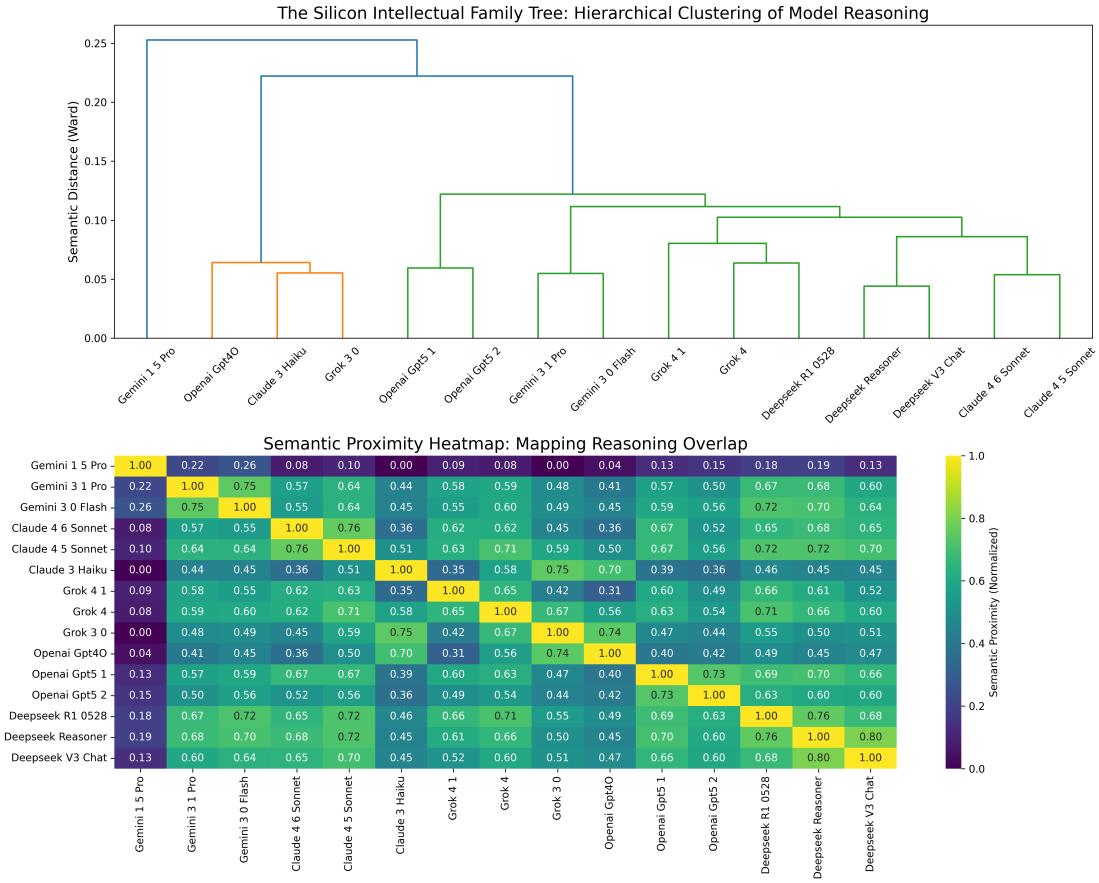
Notes: This figure compares the human expert uncertainty rate (percentage of "Uncertain" votes) against the AI semantic uncertainty mass (the probability mass assigned to the "Uncertain" bin in the SSR protocol) across the top 10 JEL categories.

Figure A.5: Individual Model Performance Dynamics by Year



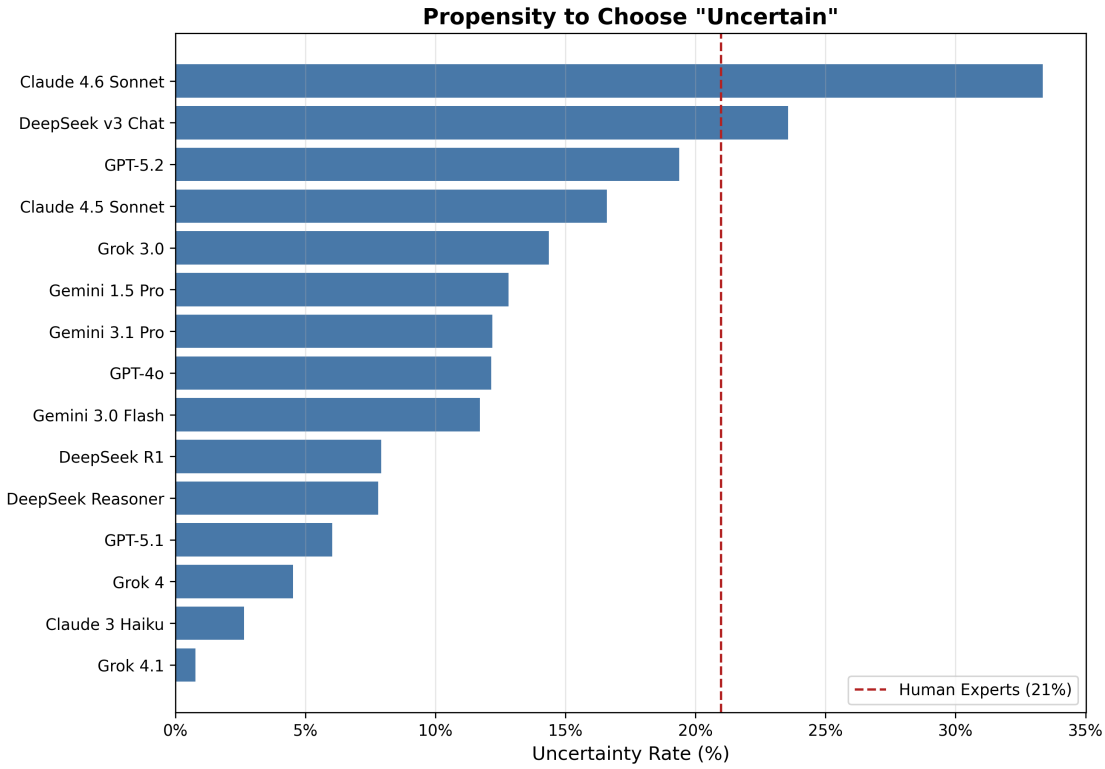
Notes: This figure provides a longitudinal view of economic alignment for each of the 15 models in our panel. Plots are color-coded by family (Google, OpenAI, Anthropic, xAI, DeepSeek) and include individual linear trend lines (dashed).

Figure A.6: The Silicon Intellectual Family Tree: Semantic Lineage and Clustering



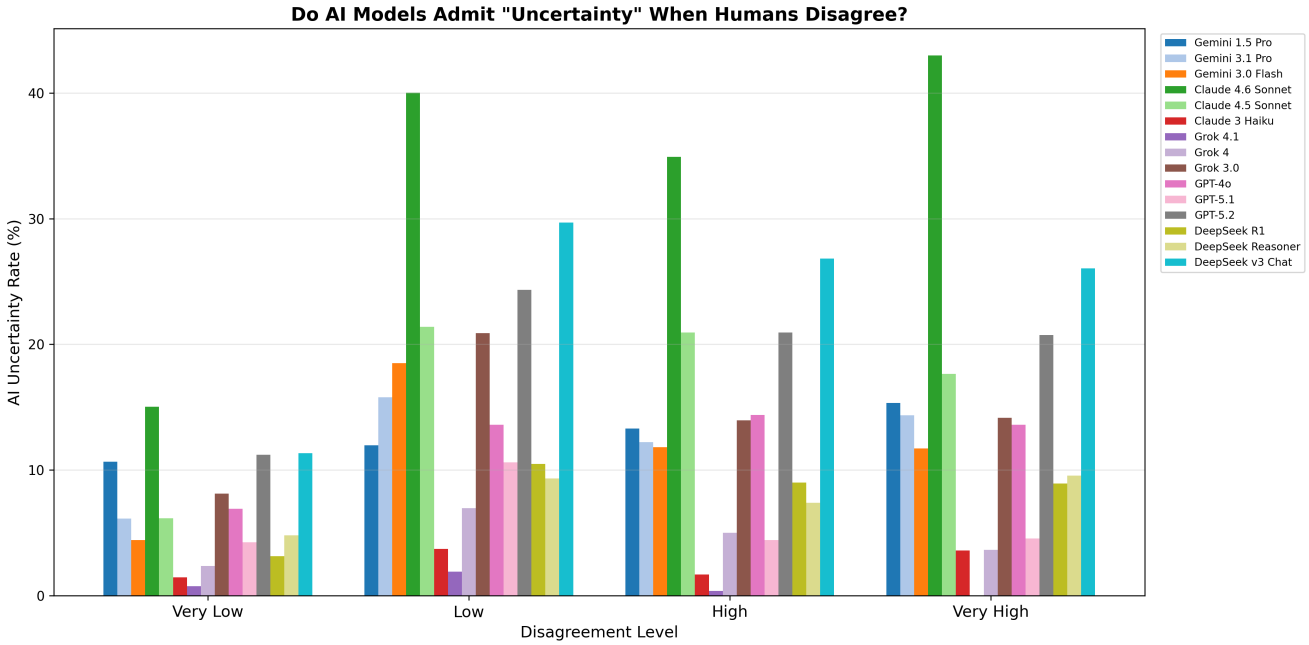
Notes: This figure visualizes the "intellectual distance" between models based on a hierarchical clustering (Ward's method) of their mean semantic reasoning embeddings across all 885 economic propositions. The bottom panel provides a heatmap of normalized semantic proximity. It reveals which architectural families share the most similar reasoning structures.

Figure A.7: Propensity to Choose "Uncertain"



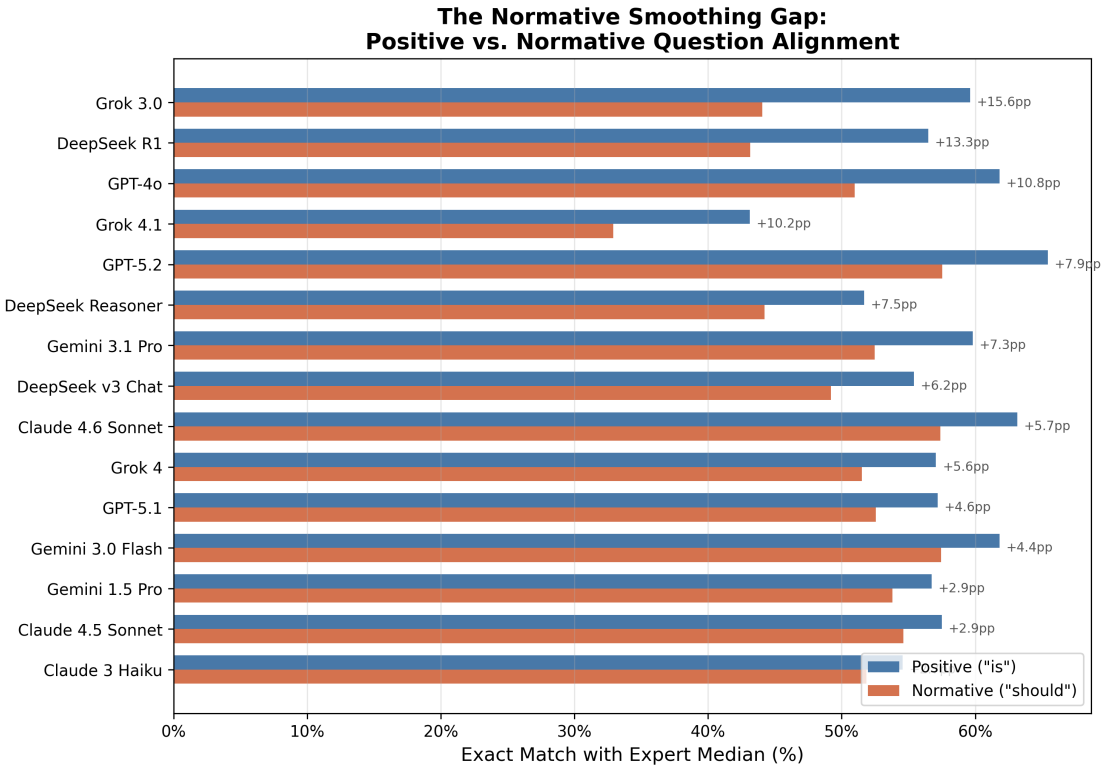
Notes: This figure ranks all 15 LLMs by the percentage of questions on which they select "Uncertain" as their vote. The dashed red line marks the human expert baseline (21%). Claude Sonnet 4.6 is the most uncertainty-prone model (34%), while Grok 4.1 almost never admits uncertainty (<1%). The 34-fold inter-model range reveals fundamentally different epistemic strategies encoded across architectures.

Figure A.8: Do AI Models Admit "Uncertainty" When Humans Disagree?



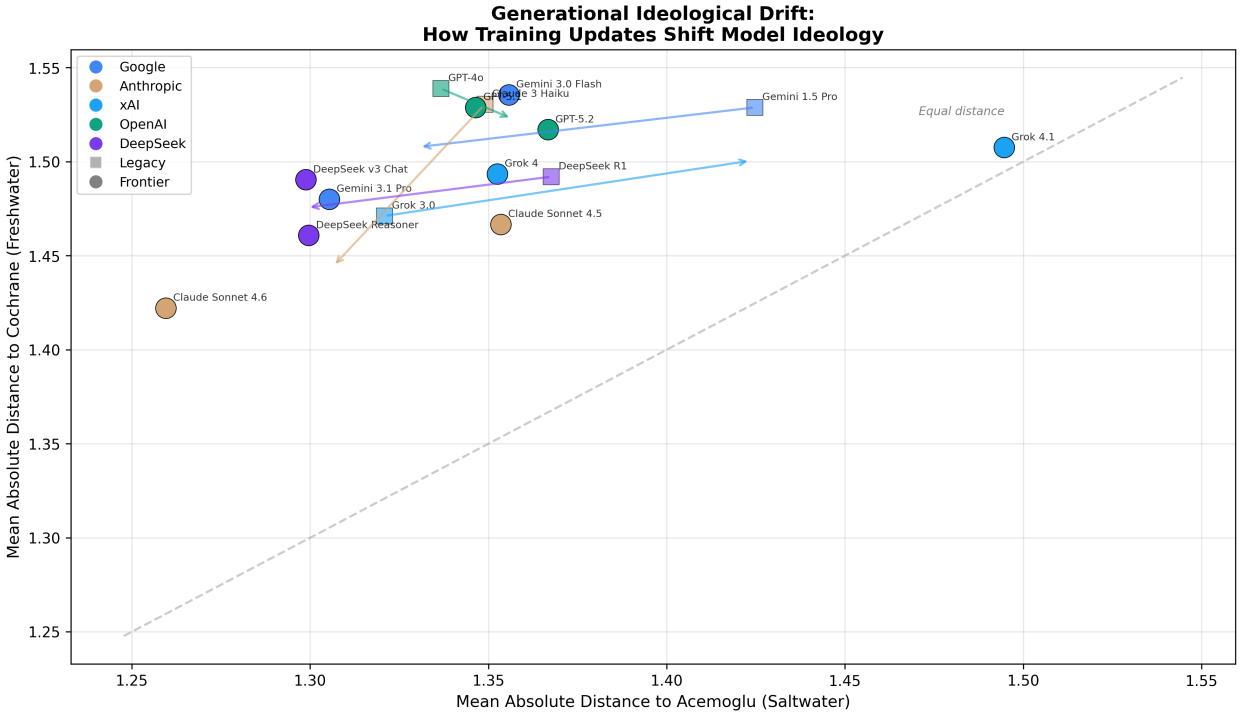
Notes: This figure shows each model's "Uncertain" vote rate across four quartiles of human expert disagreement (measured by expert vote standard deviation). Models with genuine epistemic awareness—such as Claude Sonnet 4.6 and DeepSeek v3 Chat—increase their uncertainty rate as human disagreement rises. Others—notably Grok 4.1 and Claude 3 Haiku—remain near zero regardless of question difficulty, suggesting their vote distributions are insensitive to genuine ambiguity.

Figure A.9: The Normative Smoothing Gap: Positive vs. Normative Question Alignment



Notes: This figure compares each model's exact match rate on positive ("is") vs. normative ("should") questions, classified by keyword detection in question text (134 normative, 751 positive). Every model performs worse on normative questions. The gap ranges from 2.7 pp (Claude 3 Haiku) to 15.6 pp (Grok 3.0), suggesting that value-laden economic propositions pose a systematically harder alignment challenge than descriptive ones.

Figure A.10: Generational Ideological Drift Within Architectural Families



Notes: This figure plots each model's mean absolute distance to the simulated Acemoglu anchor (x-axis) against its mean absolute distance to the simulated Cochrane anchor (y-axis) across all common questions. Points are colored by architectural family and shaped by generation (square = legacy, circle = frontier). Arrows trace the within-family shift from the legacy model to the frontier centroid. Models above the diagonal are closer to Acemoglu (Saltwater); models below are closer to Cochrane (Freshwater). Anthropic, xAI, and OpenAI shift toward Cochrane across generations, while Google and DeepSeek shift toward Acemoglu.