The Collapse of Heterogeneity in Silicon Philosophers

YUANMING SHI, Adobe Inc., USA
ANDREAS HAUPT, Stanford University, USA

Silicon samples are increasingly used as a fast and inexpensive substitute for human panels and have been shown to reproduce aggregate human opinion with high algorithmic fidelity. We show that, in the alignment-relevant domain of philosophy, silicon samples systematically collapse heterogeneity. Using data from $N = 277$ professional philosophers drawn from PhilPeople profiles, we evaluate seven proprietary and open-source large language models on their ability to replicate individual philosophical positions and to preserve cross-question correlation structures across philosophical domains. We find that language models substantially over-correlate philosophical judgments, producing artificial consensus across domains. This collapse is driven in part by specialist effects, whereby models implicitly assume that domain specialists hold highly similar philosophical views. We assess the robustness of these findings by studying the impact of DPO fine-tuning and by validating results against the full PhilPapers 2020 Survey ($N = 1{,}785$). We conclude by discussing implications for alignment, evaluation, and the use of silicon samples as substitutes for human judgment.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Philosophical/theoretical foundations of artificial intelligence*; *Knowledge representation and reasoning*.

Additional Key Words and Phrases: large language models, silicon sampling, algorithmic fidelity, expert simulation, philosophy

ACM Reference Format:
Yuanming Shi and Andreas Haupt. 2026. The Collapse of Heterogeneity in Silicon Philosophers. In *Proceedings of xxxx (XXX '26)*. ACM, New York, NY, USA, 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Silicon sampling—conditioning large language models (LLMs) on demographic and attitudinal information to generate synthetic human responses—has emerged as a promising methodology for social science research. Argyle et al. [2] demonstrated that GPT-3 exhibits substantial algorithmic fidelity when simulating U.S. political opinions, replicating not merely individual preferences but complex correlation patterns between attitudes. Park et al. [12] extended these findings through interview-based conditioning, achieving accuracy comparable to human test-retest reliability. These results suggest that LLMs capture structured belief systems rather than isolated responses, with implications for rapid, and low-cost, social science research.

However, recent work raises concerns about whether silicon samples faithfully preserve heterogeneity within human populations. Santurkar et al. [15] found that LLM opinions systematically diverge from the general U.S. population. Durmus et al. [5] showed that models struggle to faithfully represent minority viewpoints. Wang et al. [18] demonstrated that LLMs flatten identity groups, amplifying stereotypes rather than capturing authentic diversity. These findings suggest that silicon sampling may produce artificial consensus where genuine disagreement exists.

Authors' Contact Information: Yuanming Shi, jeremyshi@adobe.com, Adobe Inc., San Jose, California, USA; Andreas Haupt, h4upt@stanford.edu, Stanford University, Departments of Economics and Computer Science, Stanford, California, USA.

Manuscript submitted to ACM

Professional philosophy provides an ideal test case for whether silicon sampling preserves heterogeneity in experts-led but disagreement-rich disciplines. The PhilPapers Survey [4] documents 1,785 philosophers' positions across metaphysics, epistemology, ethics, and philosophy of mind. The survey reveals structured disagreement: philosophical positions exhibit strong correlation structures (e.g., physicalists tend toward atheism and naturalism), and variance patterns differ systematically across domains (Philosophy of Religion shows 7× more heterogeneity than Decision Theory in our data). These patterns enable comprehensive testing of whether LLMs preserve the natural diversity of expert positions or collapse it into artificial consensus.

This paper investigates whether LLMs can preserve these patterns when simulating philosophers—whether they maintain the natural heterogeneity observed in human populations and whether they replicate the correlation structure of philosophical positions. We find that LLMs systematically collapse heterogeneity: they over-correlate philosophical judgments, producing artificial consensus, and exhibit spurious specialist effects whereby models assume domain specialists hold stereotypically aligned views. We investigate two complementary dimensions of algorithmic fidelity:

RQ1 (Heterogeneity and Structural Alignment): Do silicon samples exhibit the same magnitude and structure of philosophical disagreement as human philosophers? This examines overall variance levels, domain-level predictability, and whether the latent dimensions of disagreement (identified via principal component analysis) correspond between LLMs and humans. We also assess whether LLMs produce realistic correlations between professional attributes and philosophical positions.

RQ2 (Correlation Structure Preservation): Do silicon samples preserve the empirically observed correlation structures between philosophical positions, and can fine-tuning improve this preservation? This focuses on whether LLMs maintain internal coherence when answering different questions, as human philosophers exhibit significant correlations between related positions.

These questions carry implications for AI alignment. RQ1 reveals that professional *philosophers*—those arguable best equipped to provide judgments on alignment topics—exhibit substantial disagreement; there is no expert consensus for alignment to converge upon. If LLMs collapse this heterogeneity, they risk imposing artificial consensus where genuine philosophical disagreement exists. RQ2 examines whether LLMs preserve the structured relationships between philosophical *positions* that characterize coherent worldviews. Our findings show systematic failures on both dimensions: LLMs produce 1.4–2.4× lower variance than human philosophers and organize disagreement along fundamentally different axes. While fine-tuning with Direct Preference Optimization (DPO) improves correlation structure preservation, it does not resolve the underlying heterogeneity collapse. These results suggest that current LLMs may be unsuitable for value elicitation or alignment applications that require faithful representation of philosophical diversity.

Our contributions in the following paper include: (1) the first systematic test of algorithmic fidelity in an expert domain; (2) a new evaluation framework for heterogeneity preservation using principal component analysis following Bourget and Chalmers [4]; (3) identification of systematic over-correlation and spurious specialist effects across all tested models; and (4) analysis of fine-tuning trade-offs in DPO.

2 Related Work

Silicon Sampling and Algorithmic Fidelity. Argyle et al. [2] operationalize algorithmic fidelity through four criteria: (1) *Social Science Turing Test*—generated responses are indistinguishable from human texts; (2) *Backward Continuity*—responses are consistent with the socio-demographic conditioning context; (3) *Forward Continuity*—responses proceed naturally from the context, reflecting appropriate form and tone; and (4) *Pattern Correspondence*—responses reflect underlying patterns of relationships between ideas, demographics, and behavior observed in human data. Pattern

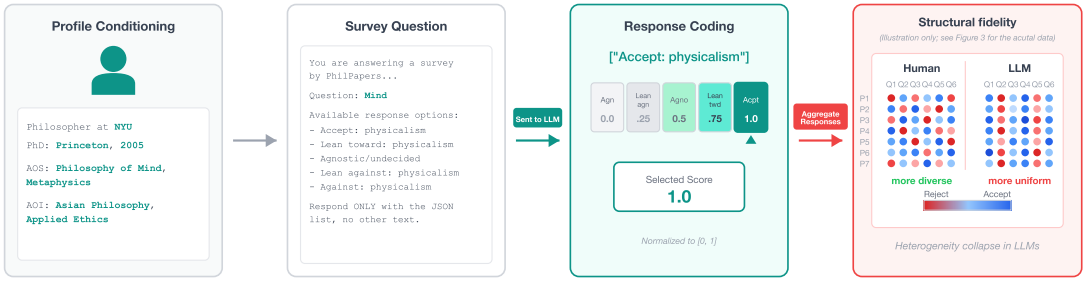


Fig. 1. Silicon sampling workflow and implication illustrated through four stages: (1) Profile Conditioning: philosopher demographics and specializations form the conditioning context; (2) Survey Question: philosophers respond to PhilPapers survey questions with discrete options; (3) Response Coding: answers are normalized to [0,1] scale; (4) Illustration of the results on Structural Fidelity Analysis. LLM responses show less heterogeneity compared to human responses.

correspondence is most relevant here, as it tests whether models capture structured belief systems rather than isolated responses. Their work on U.S. voter simulation established that LLMs can replicate complex correlations between political attitudes, demographics, and behaviors. Park et al. [12] extended this through interview-based conditioning with 1,052 individuals, achieving 85% accuracy comparable to human test-retest reliability. Beyond political opinions, Aher et al. [1] introduced "Turing Experiments," demonstrating that models reproduce known human behavioral patterns in ultimatum games and prisoner's dilemmas. Filippas et al. [6] proposed *homo silicus*—LLM-based simulated economic agents—demonstrating their utility for pilot experiments in economics. However, these investigations examined general populations on commonplace topics accessible to non-experts.

LLM Evaluation and Persona Simulation. While silicon sampling shows promise for aggregate simulation, concerns about within-group heterogeneity have emerged. Santurkar et al. [15] documented systematic liberal leanings in LLM opinions. Durmus et al. [5] showed that opinion distributions vary substantially across models and can be shifted through prompting, but minority viewpoints remain underrepresented. Wang et al. [18] provided the most systematic evidence of heterogeneity collapse, finding that LLMs reduce within-group variation and amplify stereotypes rather than capturing authentic diversity.

Philosophy as an Expert Domain. The PhilPapers Surveys [3, 4] document 1,785 professional philosophers' positions on 100 questions, providing uniquely rich data for studying expert positions. Specifically, Bourget and Chalmers [4] performed principal component analysis to characterize the dimensionality of philosophical disagreement, finding that six components each explain at least 2% of variance. Schwitzgebel and Cushman [16] documented that professional philosophers exhibit cognitive biases despite expertise, suggesting philosophical reasoning combines systematic frameworks with persistent heuristics. To our knowledge, no prior work has tested whether algorithmic fidelity extends to expert domains filled with disagreement such as philosophy.

3 Methods

Figure 2 illustrates the experimental pipeline. We explain our methods in the following sections.

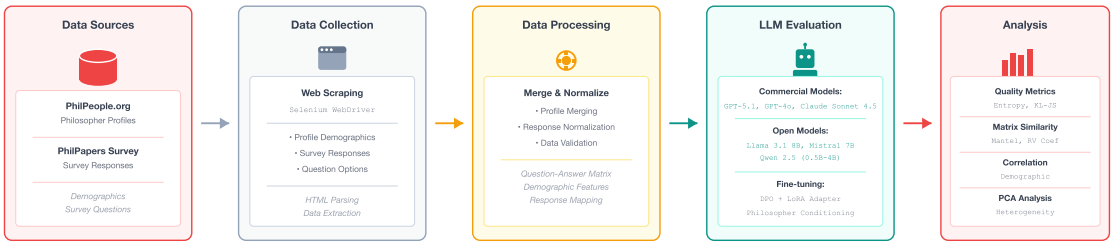


Fig. 2. Complete experimental pipeline in five stages: (1) Data Sources: philosopher profiles from PhilPeople.org and survey responses from PhilPapers; (2) Data Collection: web scraping using Selenium to extract demographics and survey data; (3) Data Processing: merging profiles with responses, normalization, and validation; (4) LLM Evaluation: conditioning both commercial models (GPT-5.1, GPT-4o, Claude Sonnet 4.5) and open-source models (Llama 3.1 8B, Mistral 7B, Qwen 3 4B), with DPO fine-tuning on 2009 survey data; (5) Analysis: quality metrics (entropy, KL-JS divergence), matrix similarity (Mantel test, RV coefficient), demographic correlation analysis, and PCA-based heterogeneity assessment.

3.1 Data and Models

Our dataset comprises 277 philosophers from PhilPapers (https://philpapers.org/) with 5,281 total survey responses. Each profile includes areas of specialization (AOS), areas of interest (AOI), PhD institution/country, graduation year, and self-reported positions (Accept, Lean toward, Agnostic, Lean against, Reject). Our sample exhibits geographic bias, overrepresenting North American philosophers (73.8% vs. 50.4% in PhilPapers 2020), but preserves topical diversity: top AOS categories match the official survey within 8 percentage points, drawn from 129 PhD institutions across 16 countries. We use the 2009 PhilPapers Survey results [3] for fine-tuning, which preserve correlation structures despite lacking explicit demographic data.

Table 1 summarizes the response statistics across all evaluated models. Following Bourget and Chalmers [4]'s methodology, we select one variable per question (highest variance for non-binary questions) and compute pairwise correlations without imputation. Human data exhibits high missingness (80.9%) as philosophers selectively answer questions in their areas of expertise; LLM data shows lower missingness (30–64%), reflecting parsing failures and refusals.

Model	N	Q	Responses	Resp%	Per-Q Var
Human	277	100	5,281	19.1%	**0.062**
GPT-4o	277	100	13,418	48.4%	0.027
GPT-5.1	276†	100	13,565	49.1%	0.026
Claude Sonnet 4.5	277	100	10,036	36.2%	0.043
Llama 3.1 8B	277	100	17,456	63.0%	0.042
Llama 3.1 8B (FT)	277	100	18,466	66.7%	0.028
Mistral 7B	277	100	19,388	70.0%	0.029
Qwen 3 4B	277	100	17,465	63.1%	0.028

Table 1. Model response statistics. N = philosophers, Q = questions, Resp% = response rate, Per-Q Var = average within-question variance. Human philosophers show 1.4–2.4× higher per-question variance than all LLMs, demonstrating systematic heterogeneity collapse. †One silicon philosopher returned empty results.

3.2 Evaluation Metrics

Evaluating algorithmic fidelity requires assessing multiple dimensions of similarity between LLM-generated and human data. We employ complementary metrics from ecology, information theory, and multivariate statistics, each capturing different aspects of fidelity.

Demographic-Position Correlations. We compute point-biserial correlations between demographic features (AOS, AOI, PhD country/year; 40 features total) and philosophical positions. Silicon sampling's core claim is that LLMs leverage demographic conditioning to generate realistic responses; systematically amplified or spurious correlations would indicate models apply stereotypes rather than learned patterns. With 9,000 tests per model, we report both uncorrected significance rates (main text) and Bonferroni-corrected rates (Appendix C) for comparison.

Specialist Effect Validation. We validate LLM specialist effects against two sources: PhilPapers 2020 Survey ($N = 1{,}785$) and our 277-philosopher ground truth. Two-source validation provides stronger evidence for spurious associations, avoiding conflation of sample-specific artifacts with genuine LLM biases. For each specialist effect, we compute chi-squared statistics with Yates correction. An effect is *spurious* if the ground truth shows no significance ($p > 0.05$) but LLMs show significant associations.

Matrix Correlation Methods. To assess whether LLMs preserve relationships between philosophical questions, we compute question-to-question correlation matrices for both human and model data, then compare these matrices. Philosophical coherence manifests in systematic relationships between positions; models that predict individual answers correctly but fail to preserve correlation structures would lack genuine philosophical reasoning. We use the Mantel test [10], a permutation-based procedure that assesses whether two correlation matrices are more similar than expected by chance (significance established via 1,000 random permutations), and the RV coefficient [14], a multivariate generalization of R^2 ranging from 0 (no similarity) to 1 (identical structure).[1]

Information-Theoretic Measures. To compare the distributions of pairwise correlations, we employ Kullback-Leibler (KL) divergence [8] and Jensen-Shannon (JS) divergence [9]. These measures capture distributional similarity beyond single correlation values, revealing whether LLMs systematically over-correlate or under-correlate questions relative to humans. We discretize correlation values into 20 bins spanning $[-1, 1]$ and compute:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

$$D_{\mathrm{JS}}(P\|Q) = \frac{1}{2} D_{\mathrm{KL}}(P\|M) + \frac{1}{2} D_{\mathrm{KL}}(Q\|M) \tag{2}$$

where $M = \frac{1}{2}(P + Q)$. Lower values indicate closer distributional match to human data.

Response Diversity. We measure response diversity using Shannon entropy [17] of the response distribution per question. Heterogeneity collapse manifests as reduced entropy: models generating near-uniform responses where humans exhibit genuine disagreement. We compute:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{3}$$

[1] The RV coefficient is defined as $\mathrm{RV}(X, Y) = \mathrm{trace}(XX'YY')/\sqrt{\mathrm{trace}(XX')^2 \cdot \mathrm{trace}(YY')^2}$.

where $p(x_i)$ is the proportion of responses taking value x_i. Higher entropy indicates more diverse responses; entropy of 0 indicates uniform responses. We track the percentage of questions with non-zero variance.

Principal Component Analysis. To characterize the dimensionality and structure of philosophical disagreement, we apply principal component analysis following Bourget and Chalmers [4]'s methodology. As described in their paper: "Our principal component analysis used only one of the numerical variables described in the preceding section for each question, so a total of 101 variables (for non-binary questions, we selected the variable with greatest variance)." We replicate this approach: for each question, we select one numerical variable (choosing the highest-variance option for non-binary questions), compute pairwise correlation matrices without imputation, and perform eigendecomposition.

PCA is appropriate because philosophical positions are not independent—beliefs form coherent worldviews with systematic correlation patterns—and PCA reveals whether LLMs preserve this structured disagreement or collapse it into artificial consensus. We avoid data imputation because mean-filling missing values fabricates positions and inflates apparent correlation strength, biasing variance estimates upward.

Despite substantial missing data in human responses (80.9% missing in our downloaded sample, as philosophers selectively answer questions in their areas of expertise), PCA with pairwise deletion is justified for three reasons. First, the missing data pattern is plausibly Missing At Random (MAR): philosophers skip questions they haven't considered deeply, not systematically based on their actual positions. Second, pairwise deletion preserves substantially more information than listwise deletion (which would discard any philosopher with any missing response). Third, our goal is descriptive—characterizing variance structure—rather than predictive, making PCA more robust to missingness than techniques requiring complete cases. LLM data has lower missingness (30–64% depending on model), creating an asymmetry we will address as limitation in Section 5.3.

Following Bourget and Chalmers [4], we count components explaining $\geq 2\%$ variance each. While PCA mathematically extracts as many components as there are variables, only components exceeding this threshold represent meaningful dimensions of disagreement rather than noise. We report: (1) variance explained by each component, (2) number of components exceeding the 2% threshold, (3) top-loading questions for each component, and (4) loading correlations between human and LLM components after optimal alignment.

3.3 Fine-tuning with Direct Preference Optimization

To investigate whether fine-tuning can improve structural fidelity, we apply Direct Preference Optimization (DPO) [13] to Llama 3.1 8B. DPO is particularly suitable for this domain because philosophical positions are matters of informed preference rather than objective truth—philosophers endorse positions based on reasoned judgment, not factual correctness. Unlike supervised fine-tuning which treats philosophical positions as ground truth labels, DPO's preference-learning framework better captures the nature of philosophical disagreement. DPO's contrastive objective fits our task of learning relative preferences between philosophical positions. Unlike Reinforcement Learning from Human Feedback (RLHF) [11], DPO directly optimizes the policy to prefer chosen over rejected responses without explicit reward modeling:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \tag{4}$$

where y_w and y_l are chosen and rejected responses.

We construct preference pairs from philosopher survey responses: for each philosopher's actual position (chosen), we generate a rejected response by inverting the stance (e.g., Accept ↔ Reject, Lean Accept ↔ Lean Reject). Training

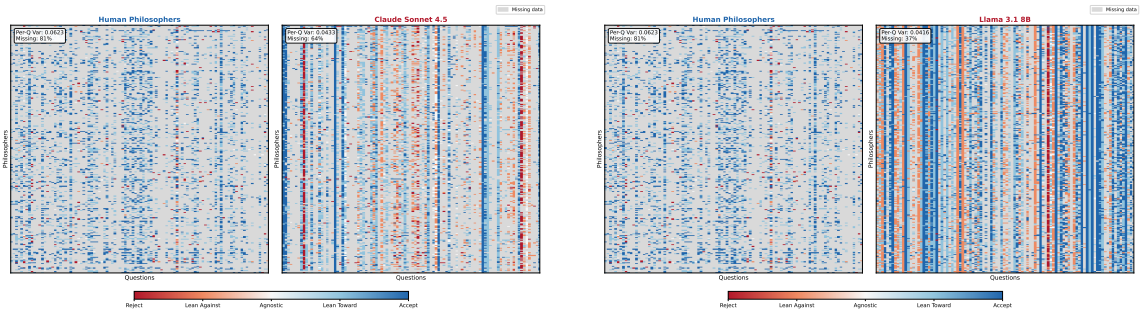
uses 3,434 examples from 226 philosophers. We apply LoRA [7] for parameter-efficient fine-tuning with rank $r = 16$, $\alpha = 32$, and dropout 0.05. Training runs for 2 epochs with batch size 1, gradient accumulation over 4 steps, learning rate 5×10^{-6}, and DPO temperature $\beta = 0.1$.

4 Results and Discussion

4.1 Experimental Setup

We evaluate seven language models: proprietary (GPT-5.1, GPT-4o, Claude Sonnet 4.5) and open-source (Llama 3.1 8B, Qwen 3 4B, Mistral 7B) systems, combining state-of-the-art and accessible smaller models. Following Argyle et al. [2], we condition each model on philosopher profiles to generate synthetic survey responses using persona and question prompts (Appendix A). We use temperature 0 to isolate model knowledge from sampling stochasticity. Responses are coded on an ordinal scale from +2 (Accept) to −2 (Reject), normalized to $[0, 1]$. Closed-source models achieved 100% parsing success; open-source models ranged from 89.4% to 99.9% (Appendix B).

Figure 3 visualizes the core finding: heterogeneity collapse in silicon sampling. Each cell represents a philosopher's response to a question, with colors ranging from red (Reject) through white (Agnostic) to blue (Accept). Human philosophers exhibit substantial within-question variation (visible as color heterogeneity within columns), while LLM simulations produce more uniform responses. See Appendix I for comparisons across all seven models.



(a) Human vs. Claude Sonnet 4.5

(b) Human vs. Llama 3.1 8B

Fig. 3. Response matrices comparing human philosophers (left panels) and LLM simulations (right panels). Rows represent individual philosophers; columns represent specific survey questions. Each row is a philosopher ($N = 277$), each column a question (100 questions). Human per-question variance is 0.062; Sonnet 4.5 shows 0.043 (1.4× lower), Llama shows 0.042 (1.5× lower). Vertical lines indicate near-uniform responses: Claude shows zero variance on 10 questions (e.g., "philosophical knowledge," "cosmological fine-tuning"); Llama on 7 questions (e.g., "statue and lump," "sleeping beauty"). Humans show lowest variance on "other minds" (var=0.002) but highest on "arguments for theism" (var=0.17). Gray indicates missing data.

4.2 Demographic-Position Correlations

We examine whether LLMs produce realistic correlations between demographic attributes and philosophical positions (RQ1). In human data, correlations are modest (e.g., Ancient Philosophy → virtue ethics: $r = 0.15$). Table 2 shows maximum observed correlations and the percentage reaching uncorrected significance ($p < 0.05$) across 9,000 tests (values near 5% indicate chance-level; Bonferroni-corrected rates in Appendix C).

LLMs produce correlations exceeding human magnitudes. The strongest human correlation is $r = 0.29$ (Metaphysics → heavyweight realism); the strongest LLM correlation is $r = 0.93$ (GPT-5.1: Philosophy of Biology → biological

Source	N	Max $\|r\|$	% Significant
PhilPapers 2020 (Human)	1,785	0.29	~5%
GPT-5.1	277	**0.93**	13.3%
GPT-4o	277	0.89	13.9%
Claude Sonnet 4.5	277	0.82	13.7%
Llama 3.1 8B	277	0.79	15.1%
Llama 3.1 8B (FT)	277	0.71	**17.8%**
Mistral 7B	277	0.79	15.9%
Qwen 3 4B	277	0.66	14.6%
Human (Our Sample)	277	0.57	~5%

Table 2. Demographic-position correlations with uncorrected significance rates ($p < 0.05$). Human philosophers show weak correlations (max 0.29 at $N = 1{,}785$; max 0.57 at $N = 277$); all LLMs show correlations 2–3× stronger. Bonferroni-corrected rates provided in Appendix C.

personal identity), a 3.2× amplification compared to the max r for humans in 2020 survey, and 63% more than max r in our sample. LLMs also show 2–3× the rate of significant correlations (13–18% vs. 5%). Detailed per-model statistics are provided in Appendix C.

We observe three patterns, validated using chi-squared tests across three data sources:

(1) Matching correlations. The PhilPapers 2020 Survey reports that 77.8% of Philosophy of Religion specialists endorse theism (+60.8 percentage points vs. non-specialists), and GPT-5.1 shows $r = 0.88$ for this association, accurately reflecting the strong human pattern.

(2) Amplified correlations. Ancient Philosophy → Aristotelian practical reason shows +26 pp in PhilPapers, only +1.9 pp in our 277-philosopher ground truth (n.s.), but +68.9 pp average across LLMs (all 7 evaluated models significant at $p < 0.001$).

(3) Spurious correlations. Some correlations appear significant in LLM predictions but are absent from human data.

Table 3 presents candidate spurious specialist effects—cases where our 277-philosopher ground truth shows no significant effect, but LLMs generate significant associations. Due to small specialist sample sizes ($n < 5$) in our ground truth, we cannot definitively validate non-significance using χ^2 tests; however, the pattern across PhilPapers 2020 Survey ($N = 1{,}785$) corroboration and large LLM effect sizes suggests systematic LLM over-prediction.

	Ground Truth ($N = 277$)		*LLM Predictions*	
Specialist Effect	Diff	Sig	Avg Diff	Sig Models
Phil. Biology → Personal identity: biological	+11.4 pp	n.s.	+43 pp	4/7***
Phil. Biology → Personal identity: psychological	+4.1 pp	n.s.	−65.7 pp	3/7***
Ancient Phil. → Practical reason: Aristotelian	+1.9 pp	n.s.	+68.9 pp	7/7***

Table 3. Candidate spurious specialist effects: Ground truth (277 philosophers) shows no significant effect, but LLMs predict highly significant associations. $^{***}p < 0.001$ for LLM predictions (stars indicate significance of LLM-side χ^2 tests, not ground truth tests). **Limitation:** Ground truth has $n < 5$ specialists, making χ^2 tests invalid; claims rely on triangulation with PhilPapers 2020 Survey and consistent LLM over-prediction patterns. LLM sample sizes are robust ($n > 100$ per model).

The Philosophy of Biology → biological personal identity effect is particularly striking: our ground truth shows a weak, non-significant association (+11.4 pp), the PhilPapers 2020 Survey reports no significant specialist effect for

this pairing, yet we observe both proprietary and open-source models (GPT-5.1, GPT-4o, Llama 3.1 8B, Claude Sonnet 4.5) predict large positive differences (average ≈43 pp across the four of seven evaluated models that surface this question, excluding the fine-tuned variant). This suggests LLMs use demographic labels as high-precision anchors for stereotypical stances, failing to recover the nuanced agnosticism or cross-domain complexity seen in human experts. The questions most predictable from demographics are listed in Appendix E.

These findings align with Santurkar et al. [15], who showed that LLM opinions systematically diverge from human populations, and Durmus et al. [5], who documented biased representations of global opinions. Our results extend these concerns to expert domains.

4.3 Question-Level Predictability Analysis

Complementing the demographic-position analysis, we examine algorithmic fidelity from the *question* perspective: which philosophical questions and domains are LLMs most and least able to predict (RQ1 continued)? This reveals where silicon sampling succeeds and where it fails, informing researchers about which philosophical topics can be reliably simulated. We compute per-question RMSE against the 277-philosopher ground truth, averaged across all models.

Question-Level Results. Table 4 shows the most and least predictable individual questions across all 100 PhilPapers questions.

Question	RMSE	Domain
Other minds	0.114	Phil. of Mind
Semantic content	0.150	Phil. of Language
Method in history	0.162	Phil. Methodology
Aim of philosophy	0.164	Phil. Methodology
Values in science	0.174	Phil. of Science

(a) Most predictable

Question	RMSE	Domain
Immortality	0.485	Phil. of Religion
Args for theism	0.452	Phil. of Religion
Politics	0.444	Political & Social
True contradictions	0.427	Logic & Formal
Capital punishment	0.402	Applied Ethics

(b) Least predictable

Table 4. Most and least predictable individual questions. RMSE ranges 4× from 0.114 to 0.485.

LLM prediction accuracy varies by a factor of 4× across questions (RMSE 0.114–0.485). The most predictable questions share characteristics: established philosophical consensus ("other minds" exist), technical/methodological grounding ("values in science"), or positions frequently discussed in training corpora. The least predictable questions involve personal belief or value-laden content—"immortality" and "arguments for theism" depend heavily on religious convictions that professional training does not determine, while "politics" and "capital punishment" reflect contentious normative debates where philosophers' positions are shaped more by personal values than disciplinary background. These patterns suggest LLMs succeed where professional consensus exists but struggle where individual conviction drives philosophical positions.

Domain-Level Results. Table 5 presents philosophical domains ranked by average prediction error. We assign each of the 100 PhilPapers questions to one of 14 domains using an LLM-assisted categorization with human validation (see Appendix J for complete assignments). Questions spanning multiple domains are assigned to their primary domain based on the question's central topic. RMSE values are computed on the normalized $[0, 1]$ response scale, enabling cross-question comparison.

Rank	Domain	RMSE	N
1	Philosophy of Science	0.193	4
2	Philosophy of Language	0.204	7
3	Aesthetics	0.204	1
4	Epistemology	0.217	10
5	Philosophy of Mind	0.219	14
6	Decision Theory	0.223	3
7	Metaphysics	0.229	17
8	Philosophical Methodology	0.229	6
9	Ethics & Moral Philosophy	0.234	10
10	History of Philosophy	0.247	4
11	Political & Social Philosophy	0.253	7
12	Applied Ethics	0.265	8
13	Logic & Formal Philosophy	0.266	5
14	Philosophy of Religion	**0.368**	4

Table 5. Domain predictability ranking. RMSE measures predictability (lower = more predictable); N = number of questions per domain.

Domain-level predictability varies by a factor of 2× (RMSE 0.193–0.368), with Philosophy of Science most predictable and Philosophy of Religion least predictable. This pattern reflects that LLMs better simulate domains with established consensus (science, language) than normatively-loaded domains requiring personal conviction (religion, applied ethics). Table 6 presents per-model RMSE for selected domains, revealing substantial model-level variation. The model-level differences may reflect varying training data composition, alignment procedures, or architectural choices, though we lack direct evidence to identify specific causal factors.

Domain	GPT-5.1	GPT-4o	Sonnet 4.5	Llama 8B	Llama 8B (FT)	Mistral 7B	Qwen 4B
Philosophy of Science	0.23	0.23	0.24	0.20	0.23	0.26	**0.19**
Philosophy of Mind	0.22	0.22	0.22	0.24	0.26	0.35	0.22
Metaphysics	0.21	0.21	0.21	0.28	0.28	0.33	0.23
Applied Ethics	0.39	0.38	**0.24**	0.54	0.25	0.36	0.42
Logic & Formal Philosophy	0.40	0.41	0.41	**0.21**	0.27	0.36	0.48
Philosophy of Religion	0.38	0.38	**0.19**	0.69	0.32	0.46	0.27

Table 6. Per-model RMSE by domain.

4.4 Principal Component Analysis: Structural Alignment

Following Bourget and Chalmers [4]'s methodology, we apply PCA to characterize the dimensionality of philosophical disagreement and assess structural alignment between human and LLM responses (RQ1 continued). Table 7 summarizes the PCA results across all sources.

Human PCA structure. Human philosophers exhibit 6 principal components each explaining ≥2% variance, with the top 6 components explaining 22.4% of total variance. PC1 (7.5%) loads most strongly on true contradictions (+0.21), philosophical progress (+0.18), and possible worlds (+0.18). PC2 (3.3%) captures meta-ethics (−0.24), ought implies can (+0.24), and epistemic justification (+0.23). PC3 (3.3%) loads on meta-ethics (−0.30), abstract objects (−0.27), and temporal ontology (−0.22). This multi-dimensionality indicates disagreement occurs along multiple independent axes without

Source	Var(6)	Elem r	Load r	Q. Overlap
Human ($N = 277$)	22.4%	—	—	—
Claude Sonnet 4.5	31.7%	0.108	0.044	0.5/5
Llama 3.1 8B (FT)	33.5%	0.079	0.083	0.5/5
Llama 3.1 8B	30.0%	0.072	0.075	0.8/5
Mistral 7B	29.1%	0.057	0.046	0.7/5
GPT-4o	28.0%	0.055	−0.022	0.5/5
GPT-5.1	28.9%	0.049	0.013	0.2/5
Qwen 3 4B	29.5%	0.016	0.031	0.5/5

Table 7. PCA structural comparison (ranked by element-wise r). Var(6) = variance explained by top 6 components; Elem r = element-wise correlation between question correlation matrices (higher is better); Load r = correlation between flattened PCA loading matrices; Q. Overlap = average top-5 question overlap per component.

any single dimension dominating—consistent with Bourget and Chalmers [4]'s finding that philosophical views do not organize along a single "left-right" spectrum. Full component interpretations are provided in Appendix H.

LLM vs. Human PC1 Comparison. Table 8 contrasts the top-5 loadings of the first principal component across models. Human PC1 spans multiple philosophical domains (logic, metaphysics, philosophy of mind), while LLM PC1s show greater topical clustering: GPT-4o and Claude Sonnet load heavily on consciousness and mind-related questions, with Claude also clustering theism questions together. This pattern supports our claim that LLMs organize disagreement along narrower, surface-topic dimensions rather than the cross-domain patterns observed in human philosophical reasoning.

Human		**GPT-4o**	
True contradictions	+0.21	Consciousness: dualism	+0.22
Philosophical progress	+0.18	Belief or credence	+0.18
Possible worlds	+0.18	Mind: non-physicalism	+0.18
Mind uploading	+0.18	Mind uploading	−0.17
Other minds	−0.17	Analysis of knowledge	+0.17
Claude Sonnet 4.5		**Llama 3.1 8B**	
Consciousness: functionalism	+0.22	Principle of sufficient reason	+0.18
Consciousness: dualism	−0.22	Material composition	+0.16
Args for theism: cosmological	−0.22	Newcomb's problem	+0.16
Mind: non-physicalism	−0.21	Extended mind	+0.16
Args for theism: ontological	−0.20	Mind: physicalism	+0.16

Table 8. PC1 top-5 loadings comparison. Human PC1 spans logic, metaphysics, and philosophy of mind. LLM PC1s cluster more narrowly by topic: GPT-4o and Claude 4.5 concentrate on consciousness/mind questions (4 of top-5), with Claude also clustering theism. Llama shows more diversity but smaller loading magnitudes.

Structural alignment metrics. We assess structural alignment using three complementary metrics: *element-wise correlation* (whether question pairs correlate similarly), *loading correlation* (whether questions have similar PCA loadings), and *question overlap* (whether top-loading questions match). Full metric definitions appear in Appendix H.

Results. All models capture more variance in their top-6 components (28–34%) than humans (22.4%), indicating LLM disagreement collapses onto fewer dimensions. Element-wise correlations are uniformly weak (0.016–0.108), as are

loading correlations (−0.02 to 0.08). Claude Sonnet 4.5 achieves the best structural alignment, followed by fine-tuned Llama. Notably, model scale does not predict alignment: GPT-5.1 underperforms smaller models. Human PC1 spans logic, metaphysics, and philosophy of mind, while LLM PC1s cluster by surface topic (e.g., theism-related questions together). Per-component comparisons in Appendix H show best individual alignments around $|r| = 0.28$, with most below 0.20.

4.5 Question Correlation Structure and the Effects from Fine-tuning

The preceding analyses examined whether LLMs correctly predict *which* positions philosophers hold and whether these predictions depend appropriately on demographic features. We now examine whether LLMs preserve the relationships *between* philosophical questions (RQ2). In human data, positions on different questions are correlated: physicalists tend toward atheism ($r \approx 0.52$) and naturalism ($r \approx 0.62$), and ethics questions cluster together. Table 9 reports metrics assessing this structural preservation.

Model	Elem r↑	RV↑	KL↓	JS↓
Llama 3.1 8B (FT)	0.083***	**0.269**	**0.062**	**0.146**
Claude Sonnet 4.5	**0.156*****	0.214	0.127	0.220
GPT-4o	0.094***	0.221	0.346	0.309
GPT-5.1	0.089***	0.214	0.362	0.322
Llama 3.1 8B	0.064***	0.257	0.155	0.200
Mistral 7B	0.041*	0.253	0.202	0.257
Qwen 3 4B	0.041	0.200	0.461	0.351

Table 9. Question correlation structure preservation. ↑/↓ indicates higher/lower is better. $^{***}p < 0.001$, $^{*}p < 0.05$ (Mantel test).

Different metrics favor different models. Claude Sonnet 4.5 best preserves specific question-pair relationships (element-wise $r = 0.156$), while the fine-tuned model best preserves the overall distribution of correlations (KL = 0.062, 60% better than base model). Notably, GPT-5.1 underperforms GPT-4o on structure preservation despite stronger overall capabilities, suggesting structural fidelity is not a simple function of model scale. Full model comparison statistics are provided in Appendix F. Aggregating across all metrics, Claude Sonnet 4.5 achieves the best overall performance, followed by Llama 3.1 8B (best open-source). GPT-5.1 ranks third despite having the strongest demographic correlations, penalized by poor structural preservation.

In addition, DPO fine-tuning produces mixed effects on different dimensions of fidelity. Table 10 compares base and fine-tuned Llama 3.1 8B across response diversity and structural coherence metrics.

	Response Diversity		*Structural Coherence*	
Model	Entropy↑	Resp. KL↓	Corr. KL↓	Corr. JS↓
Llama 3.1 8B (Base)	0.662	5.53	0.155	0.200
Llama 3.1 8B (FT)	0.605	8.23	0.062	0.146

Table 10. Fine-tuning trade-off. DPO improves structural coherence (KL −60%) but reduces response diversity (entropy −8.6%).

The fine-tuned model collapses 15 questions to uniform responses (all philosophers receive identical answers), yet better preserves correlations between questions. Multiple mechanisms could explain this trade-off: DPO training may

have overfit to the training distribution, the limited training data (3,434 examples) may be insufficient to capture both structural relationships and within-question variation, or the contrastive DPO objective may bias toward learning point estimates of "correct" positions rather than preserving response distributions. Distinguishing these mechanisms requires further investigation with varying data scales and training objectives. Per-question effects are detailed in Appendix G.

4.6 Implications for AI Alignment

These findings raise concerns for using silicon samples as proxies for expert judgment. LLMs organize philosophical disagreement along fundamentally different dimensions than humans, clustering questions by surface-level topic similarity rather than the cross-domain patterns observed in human data. This suggests models rely on lexical associations rather than coherent philosophical reasoning—a pattern consistent across all seven models, indicating systematic LLM properties rather than model-specific artifacts.

Amplified correlations and spurious specialist effects further demonstrate that models apply stereotypes rather than learned patterns. Our findings align with Wang et al. [18]'s observation that LLMs flatten identity groups, here extended to professional philosophers. Domains critical for normative alignment (Philosophy of Religion, Applied Ethics) show the most severe distortion.

These results suggest that silicon sampling, while useful for simulating aggregate population trends [2], may be inadequate for applications requiring faithful representation of expert disagreement. Researchers should exercise particular caution when using LLMs to simulate expert judgment in normatively-loaded domains, and explicitly measure structural alignment rather than only aggregate accuracy.

5 Conclusion, Future Work and Limitations

5.1 Conclusion

We evaluated seven LLMs on their ability to replicate philosopher survey responses, comparing against the PhilPapers 2020 Survey and our 277-philosopher ground truth. We identified three key findings.

First, all LLMs exhibit structural mismatch in how philosophical disagreement is organized. LLMs cluster questions by surface-level topic similarity rather than the cross-domain patterns observed in human data, capturing more variance in fewer principal components—consistent with heterogeneity collapse.

Second, LLMs exhibit inflated demographic-position correlations with spurious specialist effects. This "stereotyping" behavior, where models use demographic labels as high-precision anchors for stereotypical stances rather than recovering nuanced agnosticism or cross-domain complexity, explains their failure to capture the structures found in expert data.

Third, DPO fine-tuning improves correlation structure preservation but causes variance collapse and reduces response diversity—suggesting preference optimization faces inherent trade-offs between structural fidelity and heterogeneity preservation.

These results suggest that silicon sampling is inadequate for applications requiring faithful representation of expert disagreement in normatively-loaded domains. Alignment research must account for the fact that LLMs may produce artificial consensus that does not reflect the reasoned disagreement of experts.

5.2 Future Work

Several directions follow. First, validating against official PhilPapers individual data would strengthen claims about ground-truth patterns. Second, characterizing the temperature-fidelity trade-off would clarify whether sampling diversity mitigates heterogeneity collapse. Third, exploring alternative fine-tuning approaches (diversity-regularized objectives, varying DPO β) could determine whether the diversity-structure trade-off is inherent or can be overcome. Finally, developing metrics that explicitly quantify heterogeneity preservation alongside accuracy could improve silicon sampling evaluation.

5.3 Limitations

Sample Demographics. Our 277-philosopher sample exhibits geographic bias, overrepresenting North American philosophers (73.8% vs. 50.4% in PhilPapers 2020), with 22.7% lacking PhD country information. While topical diversity (AOS categories) is preserved, these geographic skews may affect generalizability, particularly for questions where regional philosophical traditions differ substantively.

Selection Bias. Philosophers maintaining active PhilPeople profiles may differ systematically (more online engagement, younger, more technologically literate). Self-reported positions may have changed since data collection, introducing temporal inconsistency.

Missing Data Handling. Our principal component analysis uses pairwise correlations without imputation, which represents a key methodological difference between human and LLM data analysis. While this approach avoids creating artificial philosophers and inflating correlation estimates, it may underestimate variance explained if missingness is non-random. LLM data has minimal missingness (parsing failures only), whereas human data has substantial non-response (philosophers selectively skip questions). This asymmetry complicates direct PCA comparisons, though our primary findings (heterogeneity collapse, failure modes) remain robust to this limitation.

6 Generative AI Usage Statement

This research leveraged Generative AI tools (Claude, ChatGPT) in several capacities. For code generation, AI assisted with implementing web scrapers (e.g., Selenium-based data collection from PhilPeople), data processing pipelines, and SVG-graphics generation; the authors carefully verified all generated code, conducted debugging, and took full responsibility for its correctness (or lack thereof). AI tools significantly accelerated prototyping and helped implement techniques the authors were initially unfamiliar with.

For writing, AI helped identify grammatical errors, improve sentence structure, and polish academic tone to meet publication standards.

For methodology, during the brainstorming stage, AI suggested relevant but less commonly known statistical metrics (e.g., the Mantel test for matrix correlation, RV coefficient) that proved valuable for our analysis framework. These suggestions were independently verified against statistical literature before adoption.

The authors maintain full ownership of research design, hypothesis formulation, experimental execution, result interpretation, and all conclusions drawn.

References

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 337–371.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351.
[3] David Bourget and David J Chalmers. 2014. What Do Philosophers Believe? *Philosophical Studies* 170, 3 (2014), 465–500.
[4] David Bourget and David J Chalmers. 2023. Philosophers on Philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint* 23, 11 (2023), 1–66.
[5] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388* (2023).
[6] Apostolos Filippas, John J Horton, and Benjamin S Manning. 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*. ACM.
[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021).
[8] Solomon Kullback and Richard A Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86.
[9] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.
[10] Nathan Mantel. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research* 27, 2 (1967), 209–220.
[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[12] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109* (2024).
[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36 (2023).
[14] Paul Robert and Yves Escoufier. 1976. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 25, 3 (1976), 257–265.
[15] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 29971–30004.
[16] Eric Schwitzgebel and Fiery Cushman. 2012. Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers. *Mind & Language* 27, 2 (2012), 135–153.
[17] Claude E Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.
[18] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large Language Models that Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence* 7 (2025), 400–411.

A Prompt Templates

A.1 Persona Prompt Construction

The persona prompt encodes the philosopher's professional identity:

```
You are a professional philosopher at {institution} ({country}).

Your Educational Background:
- PhD from {phd_institution} ({phd_country}) in {phd_year}

Your Areas of Specialization:
- {specialization_1}
- {specialization_2}
[up to 5 specializations]

Your Areas of Interest:
- {interest_1}
- {interest_2}
```

```
[up to 8 interests]
```

A.2 Question Prompt

```
You are answering a survey by the reputable website PhilPapers,
which collects responses across various philosophical domains
based on your background.

Please respond with your chosen option(s) as a JSON list.
If selecting multiple options, ensure they are logically consistent.

Examples of valid responses:
- ["Accept: physicalism"]
- ["Accept: physicalism", "Reject: non-physicalism"]
- ["Lean towards: physicalism", "Lean against: non-physicalism"]

Given your philosophical profile above, please express your stance
on the following question.

Question: {question_key}

Available response options:
- {option_1}
- {option_2}
...

Respond ONLY with the JSON list, no other text.
```

B Response Parsing Details

Table 11 reports parsing success rates.

Model	Success Rate	Primary Failure Mode
GPT-5.1	100.0%	—
GPT-4o	100.0%	—
Claude Sonnet 4.5	100.0%	—
Qwen 3 4B	99.9%	Invalid option format
Mistral 7B	98.1%	Invalid option format
Llama 3.1 8B	93.9%	"Combination of views"
Llama 3.1 8B (FT)	89.4%	"Combination of views"

Table 11. Response parsing success rates.

Analysis of Llama 3.1 8B failures reveals that the majority (66.3%) involve generating hedging responses not in the valid option set. DPO fine-tuning increased failures from 6.1% to 10.6%, with "combination of views" failures rising from 686 to 2,312.

C Detailed Demographic Correlation Statistics

Model	Max $\|r\|$	Mean $\|r\|$	% Sig
Human	0.571	0.099	0.01%
GPT-5.1	0.933	0.085	0.93%
GPT-4o	0.888	0.086	0.78%
Claude Sonnet 4.5	0.818	0.091	1.05%
Llama 3.1 8B	0.788	0.082	1.17%
Llama 3.1 8B (FT)	0.706	0.086	1.62%
Mistral 7B	0.789	0.082	1.63%
Qwen 3 4B	0.660	0.077	1.07%

Table 12. Detailed demographic correlation statistics with Bonferroni-corrected significance rates ($N = 277$ philosophers, $\alpha = 0.05/9000 \approx 5.6 \times 10^{-6}$).

D Demographic Feature Categories

Demographic features used in correlation analyses include four categories:

Areas of Specialization (AOS) and Areas of Interest (AOI): Ancient Greek and Roman Philosophy, 17th/18th Century Philosophy, 19th Century Philosophy, 20th Century Philosophy, Applied Ethics, Epistemology, General Philosophy of Science, Logic and Philosophy of Logic, Meta-Ethics, Metaphysics, Normative Ethics, Philosophy of Action, Philosophy of Biology, Philosophy of Cognitive Science, Philosophy of Language, Philosophy of Mind, Philosophy of Physical Science, Philosophy of Religion, Social and Political Philosophy, and others.

PhD Country: USA, United Kingdom, Canada, Germany, Australia, and others (including "Unknown" for missing data).

PhD Year: Binned by 5-year intervals (e.g., 1990–1994, 1995–1999, 2000–2004).

In total, approximately 40 unique demographic features are tested per model, yielding ~9,000 feature-question pairs.

E Most Predictable Questions from Demographics

This appendix presents the top-6 most predictable questions from demographic features for each model, showing the strongest demographic predictor and absolute correlation value.

F Full Model Comparison for Question Correlations

This appendix presents comprehensive metrics for assessing how well each model preserves the correlation structure between philosophical questions observed in human data. We report five complementary metrics: element-wise correlation (Elem r) measures agreement on specific question pairs, Mantel p tests statistical significance, RV coefficient captures overall structural similarity, and KL/JS divergence quantify distributional differences. Together, these metrics provide a multi-faceted view of structural fidelity.

Question	Strongest Predictor	$\|r\|$
GPT-5.1		
Personal identity: biological view	AOS: Phil. of Biology	0.933
Foundations of mathematics: logicism	AOS: Logic	0.891
God: atheism	AOS: Phil. of Religion	0.880
God: theism	AOS: Phil. of Religion	0.875
Quantum mechanics: hidden-variables	AOS: 17th/18th C. Phil.	0.703
Cosmological fine-tuning: design	AOS: Phil. of Religion	0.702
GPT-4o		
God: theism	AOS: Phil. of Religion	0.888
Personal identity: biological view	AOS: Phil. of Biology	0.781
Personal identity: biological view	AOI: Phil. of Biology	0.721
Foundations of mathematics: logicism	AOS: Logic	0.718
Vagueness: metaphysical	AOS: Metaphysics	0.693
Cosmological fine-tuning: design	AOS: Phil. of Religion	0.682
Claude Sonnet 4.5		
Foundations of mathematics: logicism	AOS: Logic	0.818
Cosmological fine-tuning: design	AOS: Phil. of Religion	0.806
Morality: expressivism	AOS: Phil. of Language	0.788
Free will: libertarianism	AOS: Phil. of Religion	0.709
Moral motivation: externalism	AOS: Meta-Ethics	0.685
Morality: non-naturalism	AOS: Phil. of Religion	0.679

Table 13. Most predictable questions (part 1 of 2). Models show strong correlations between specializations and lexically-related positions (e.g., Philosophy of Biology → biological view, $r > 0.7$), many spurious.

G Fine-tuning Effects by Question

This appendix presents detailed question-by-question analysis of how DPO fine-tuning affects response diversity. We compare Shannon entropy (H) before and after fine-tuning for Llama 3.1 8B. Fine-tuning produces heterogeneous effects: some questions show dramatically increased diversity (Table 16), while others collapse to near-uniform responses (Table 17). This pattern reveals the trade-off between structural coherence (improved) and response diversity (degraded) discussed in Section 4.4.

H Principal Component Analysis Details

This appendix presents detailed PCA results following Bourget and Chalmers [4]'s methodology: pairwise correlation matrix (no imputation), eigendecomposition, and 2% variance threshold for significant components. The human sample ($N = 277$) has 22.4% variance explained by the top 6 components.

H.1 Human Principal Components

Table 18 presents the top loadings for each component. Here is the breakdown with the concentrated topics in each component.

PC1 (7.5% variance): Logic and Metaphysics. Loads on true contradictions (−0.20), logic (−0.18), possible worlds (−0.18), analytic-synthetic distinction (−0.17), and metaontology (−0.17).

PC2 (3.3% variance): Ethics and Epistemology. Loads on meta-ethics (−0.24), ought implies can (+0.24), epistemic justification (+0.23), time (−0.22), and political philosophy (−0.21).

Question	Strongest Predictor	$\lvert r \rvert$
Llama 3.1 8B		
Arguments for theism: cosmological	AOS: Phil. of Religion	0.788
Personal identity: biological view	AOI: Phil. of Biology	0.780
Aesthetic experience: sui generis	AOS: Aesthetics	0.685
Practical reason: Humean	AOS: 17th/18th C. Phil.	0.675
Political philosophy: libertarianism	AOI: Social & Political Phil.	0.576
Time: A-theory	AOI: Phil. of Mind	0.566
Mistral 7B		
Gender: psychological	AOI: Phil. of Gender	0.789
External-world skepticism: externalist	AOS: Phil. of Language	0.784
Gender: social	AOI: Phil. of Gender	0.781
Normative ethics: consequentialism	AOS: Normative Ethics	0.758
Political philosophy: communitarianism	AOS: Social & Political Phil.	0.701
Proper names: Millian	AOS: Phil. of Language	0.691
Qwen 3 4B		
Moral judgment: cognitivism	AOS: Normative Ethics	0.660
Political philosophy: libertarianism	AOI: Phil. of Physical Science	0.637
Political philosophy: libertarianism	AOS: General Phil. of Science	0.616
Practical reason: Humean	AOI: Ancient Greek Phil.	0.569
Arguments for theism: design	AOS: Phil. of Religion	0.537
Arguments for theism: design	AOI: Phil. of Religion	0.529

Table 14. Most predictable questions (part 2 of 2). All models exhibit inflated demographic correlations, with proprietary models (GPT-5.1, GPT-4o, Claude Sonnet 4.5) showing stronger effects than open-source models (Llama, Mistral, Qwen).

Model	Elem r	Mantel p	RV	KL	JS
Llama 3.1 8B (FT)	0.083	<0.001	0.269	0.062	0.146
Claude Sonnet 4.5	0.156	<0.001	0.214	0.127	0.220
GPT-4o	0.094	<0.001	0.221	0.346	0.309
GPT-5.1	0.089	<0.001	0.214	0.362	0.322
Llama 3.1 8B	0.064	<0.001	0.257	0.155	0.200
Mistral 7B	0.041	0.036	0.253	0.202	0.257
Qwen 3 4B	0.041	0.086	0.200	0.461	0.351

Table 15. Full question correlation structure metrics. Element-wise r measures pairwise agreement; Mantel p tests significance via permutation; RV coefficient measures overall structural similarity (0–1); KL/JS divergence measure distributional difference (lower is better). All significance tests use $\alpha = 0.05$.

Question	Base	FT	ΔH	Δ%
Analytic-synthetic	0.53	1.42	+0.90	+170
Consciousness: func.	0.21	1.03	+0.83	+404
Continuum: indet.	0.22	0.98	+0.76	+352
Aesthetic: percept.	0.46	1.22	+0.75	+162
Env. ethics	0.11	0.71	+0.60	+553

Table 16. Questions where fine-tuning increased diversity (top 5 by absolute ΔH). H = Shannon entropy; ΔH = change in entropy; Δ% = percentage change.

PC3 (3.3% variance): Metaethics and Ontology. Loads on meta-ethics (−0.30), abstract objects (−0.27), temporal ontology (−0.22), race categories (+0.21), and morality (−0.20).

Question	Base	FT	ΔH	Δ%
Conscious.: identity	0.69	0.00	−0.69	−100
Conscious.: panpsy.	0.99	0.38	−0.61	−62
Abstract: Platonism	0.85	0.39	−0.46	−54
A priori: no	1.23	0.81	−0.42	−34
Abortion: permiss.	1.19	0.71	−0.48	−41

Table 17. Questions where fine-tuning decreased diversity (top 5 by absolute ΔH). The "consciousness: identity" question collapsed to uniform responses (H = 0), indicating all philosophers received identical predictions after fine-tuning.

PC4 (2.9% variance): Normative Theory. Loads on normative ethics (+0.36), political philosophy (+0.27/+0.27), normative ethics (+0.24), and temporal ontology (−0.20).

PC5 (2.8% variance): Intentionality and Skepticism. Loads on grounds of intentionality (+0.28), external world (+0.23), theory of reference (+0.23), time (−0.20), and meta-ethics (−0.19).

PC6 (2.5% variance): Free Will and Knowledge. Loads on free will (−0.30), epistemic justification (−0.27), knowledge (−0.25), metaphilosophy (+0.24), and logic (−0.23).

H.2 Per-Model Component Comparison

Table 19 shows per-component loading correlations (direct, not optimally aligned) between each model and human data. Most correlations are weak ($|r| < 0.2$), with best alignments for Llama PC1 (0.28) and PC3 (0.27).

Key finding: Human components span multiple philosophical domains. LLM components tend to cluster by surface topic similarity rather than cross-domain patterns.

These results reveal that philosophical disagreement is multi-dimensional, with no single dimension dominating (PC1 explains 7.5%). The top 6 components together explain 22.4% of variance. LLM responses are more compressible (28–34% variance in top-6), consistent with heterogeneity collapse.

I Heterogeneity Collapse Across All Models

Figure 4 presents response matrices for all eight data sources (human + 7 LLMs), enabling direct visual comparison of heterogeneity patterns. Each panel shows the same 277 philosophers × 100 questions matrix, with color indicating response (red = Reject, white = Agnostic, blue = Accept) and gray indicating missing data.

The visualization reveals systematic patterns: (1) Human data shows substantial color variation within columns, indicating genuine disagreement on each question; (2) All LLMs show more uniform within-column coloring, with Claude Sonnet 4.5 most closely approximating human heterogeneity; (3) Fine-tuned Llama shows notably increased "Agnostic" responses (white band) compared to base Llama; (4) Mistral 7B exhibits heavy Agnostic bias (53.9% of responses); (5) Qwen 3 4B shows the most extreme collapse, with "Lean Toward" dominating (54.5% of responses).

J Question-Domain Assignments

Table 20 presents the complete mapping of all 100 PhilPapers questions to their assigned philosophical domains. We used an LLM (Claude, distinct from the evaluation models) to initially suggest classifications based on standard philosophy curriculum boundaries, then validated and corrected all 100 assignments to ensure accuracy. Domains are ordered by predictability (RMSE), from most predictable (Philosophy of Science) to least predictable (Philosophy of Religion).

PC	Top Questions (by absolute loading)	Loading
PC1	True contradictions	−0.20
	Logic	−0.18
	Possible worlds	−0.18
	Analytic-synthetic distinction	−0.17
	Metaontology	−0.17
PC2	Meta-ethics	−0.24
	Ought implies can	+0.24
	Epistemic justification	+0.23
	Time	−0.22
	Political philosophy	−0.21
PC3	Meta-ethics	−0.30
	Abstract objects	−0.27
	Temporal ontology	−0.22
	Race categories	+0.21
	Morality	−0.20
PC4	Normative ethics	+0.36
	Political philosophy	+0.27
	Political philosophy	+0.27
	Normative ethics	+0.24
	Temporal ontology	−0.20
PC5	Grounds of intentionality	+0.28
	External world	+0.23
	Theory of reference	+0.23
	Time	−0.20
	Meta-ethics	−0.19
PC6	Free will	−0.30
	Epistemic justification	−0.27
	Knowledge	−0.25
	Metaphilosophy	+0.24
	Logic	−0.23

Table 18. Top loadings for each principal component in human data ($N = 277$). These loadings reveal the philosophical dimensions along which disagreement occurs.

Model	PC1	PC2	PC3	PC4	PC5	PC6
Llama 3.1 8B	0.28	0.10	0.27	0.05	0.13	0.01
Llama 3.1 8B (FT)	0.22	0.00	0.19	0.08	0.04	0.25
Claude Sonnet 4.5	0.01	0.11	0.21	0.02	0.22	0.15
GPT-4o	0.03	0.12	0.12	0.19	0.12	0.11
GPT-5.1	0.02	0.05	0.13	0.05	0.16	0.17
Mistral 7B	0.17	0.09	0.14	0.06	0.21	0.13
Qwen 3 4B	0.14	0.20	0.06	0.07	0.14	0.05

Table 19. Per-component loading correlations ($|r|$) between human and each model (direct comparison, not optimally aligned). Most correlations are weak ($|r| < 0.2$).

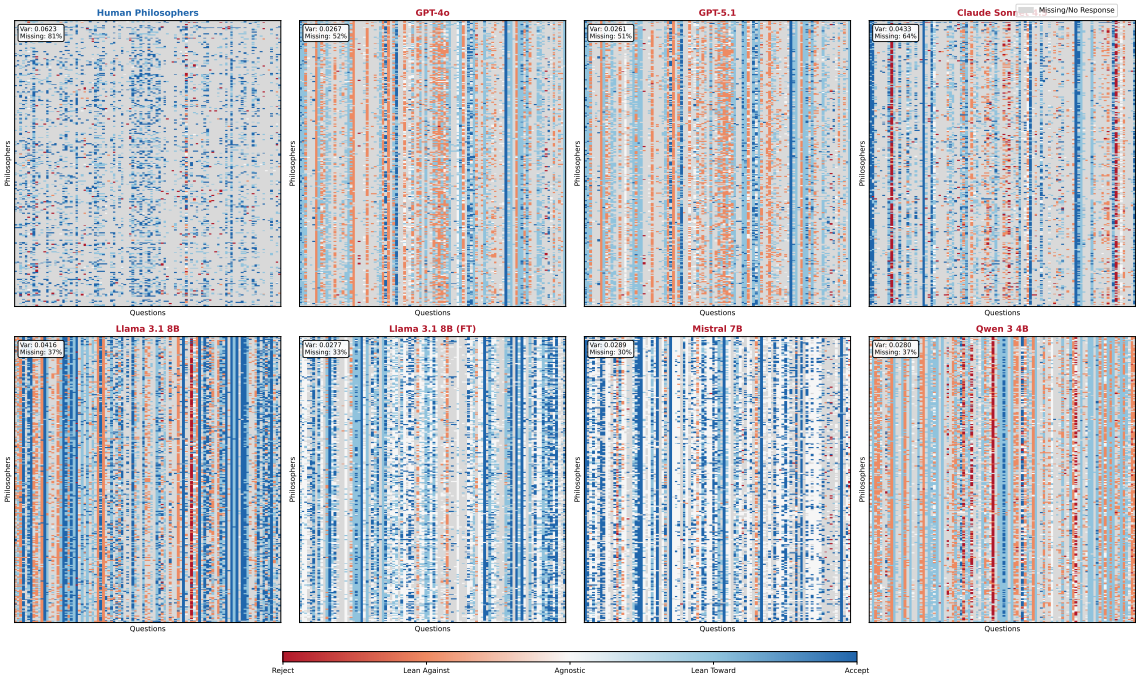


Fig. 4. Response matrices for human philosophers and all seven LLM simulations. Each panel shows 277 philosophers (rows) × 100 questions (columns). Human per-question variance (0.062) exceeds all LLMs (0.026–0.043), demonstrating systematic heterogeneity collapse across model architectures and scales.

Domain (N)	Questions
Philosophy of Science (4)	quantum mechanics, science, units of selection, values in science
Philosophy of Language (7)	analytic-synthetic distinction, proper names, propositions, semantic content, theory of reference, truth, vagueness
Aesthetics (1)	aesthetic value
Epistemology (10)	a priori knowledge, analysis of knowledge, belief or credence, epistemic justification, justification, knowledge, knowledge claims, philosophical knowledge, rational disagreement, response to external-world skepticism
Philosophy of Mind (14)	aesthetic experience, chinese room, concepts, consciousness, extended mind, grounds of intentionality, hard problem of consciousness, mental content, mind, mind uploading, other minds, perceptual experience, propositional attitudes, zombies
Decision Theory (3)	newcomb's problem, practical reason, sleeping beauty
Metaphysics (17)	abstract objects, causation, external world, free will, interlevel metaphysics, laws of nature, material composition, metaontology, personal identity, possible worlds, properties, spacetime, statue and lump, teletransporter, temporal ontology, time, time travel
Phil. Methodology (6)	aim of philosophy, metaphilosophy, method in history of philosophy, method in political philosophy, philosophical methods, philosophical progress
Ethics & Moral Phil. (10)	meaning of life, meta-ethics, moral judgment, moral motivation, moral principles, morality, normative concepts, normative ethics, ought implies can, well-being
History of Philosophy (4)	hume, kant, plato, wittgenstein
Political & Social Phil. (7)	gender, gender categories, law, political philosophy, politics, race, race categories
Applied Ethics (8)	abortion, capital punishment, eating animals and animal products, environmental ethics, experience machine, footbridge, human genetic engineering, trolley problem
Logic & Formal Phil. (5)	continuum hypothesis, foundations of mathematics, logic, principle of sufficient reason, true contradictions
Philosophy of Religion (4)	arguments for theism, cosmological fine-tuning, god, immortality

Table 20. Complete question-domain assignments. All 100 PhilPapers questions are assigned to one of 14 philosophical domains using LLM-assisted categorization (Claude Opus 4.5) with human validation.