

Simulated philosophers disagree less, and differently than real ones.

The Collapse of Heterogeneity in Silicon Philosophers

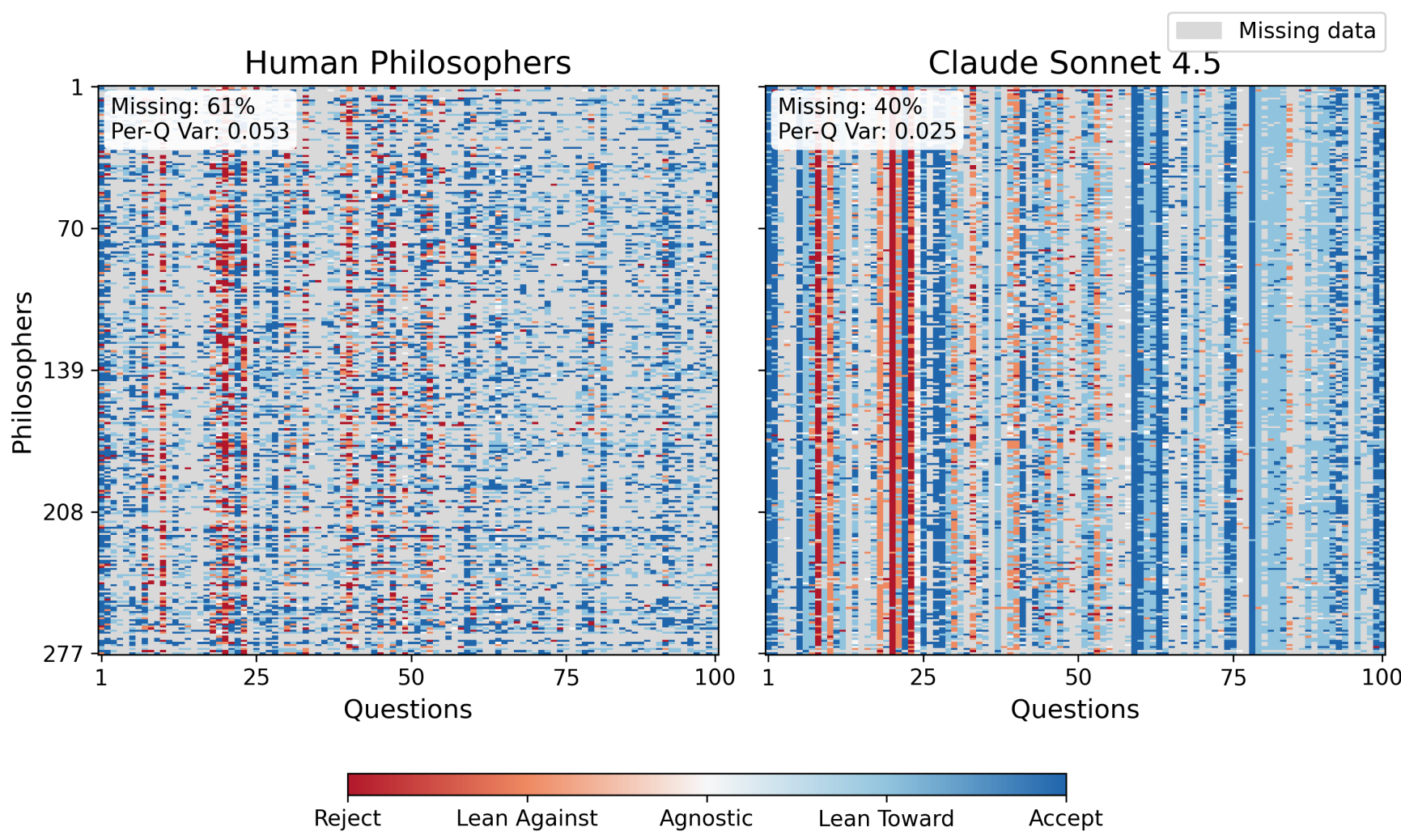
Yuanming Shi¹ Andreas Haupt²
¹Adobe Inc. ²Stanford University, Economics & Computer Science
ACM FAccT 2026, Montreal, QC

Setup

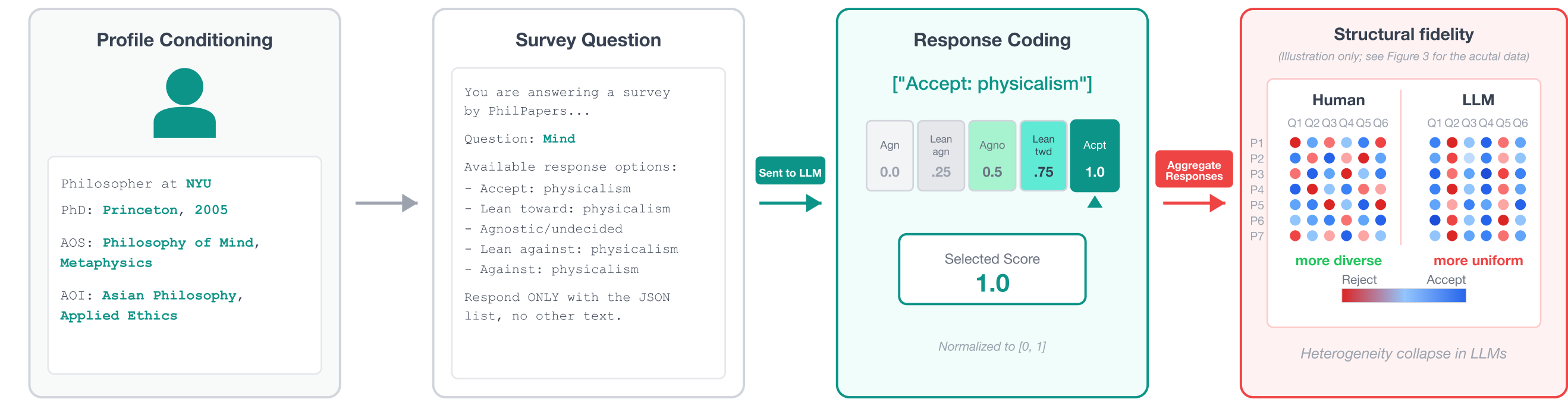
- **Ground truth:** 277 philosophers from PhilPeople and their answers to the 100 PhilPapers questions.
- **Silicon samples:** 7 models conditioned on each philosopher’s specializations, PhD school and year.
- Answers coded -2 to $+2$ at temperature 0
- Evaluated against PhilPapers 2020 ($N = 1785$).

Measurement

Not correctness but **does the structure of disagreement survive?** Variance, correlation structure, principal axes, specialist effects.



One cell per answer; rows are the same 277 philosophers, columns the same 100 questions. Humans speckled; Claude-Sonnet-4.5 — *closest* model to humans — striped.



Data

Source	Per-Q	Var(6)	Elem r	Load r
Human	.053	71.1%	—	—
Sonnet-4.5	.026	72.3%	.245**	.023
GPT-4o	.014	83.0%	.103**	.110
GPT-5.1	.014	81.6%	.099**	-.137
Mistral-7B	.028	73.8%	.027	-.041
Qwen-3-4B	.016	67.8%	.027	.045
Llama-3.1-8B	.026	62.4%	.016	.026
Llama-8B DPO	.019	61.2%	.065*	-.063

Per-Q = within-question variance; Var(6) = variance in the top 6 principal components. Proprietary models pack *more* of it into fewer axes than humans, with loadings uncorrelated to theirs.

Spurious specialist effects

Specialist → position	Human	LLM
Ancient Phil. → Aristotelian prac. reason	+1.9 n.s.	+68.9 7/7***
Phil. Biology → biological identity	+11.4 n.s.	+43.0 4/7***
Metaphysics → Humean laws	-8.2 n.s.	-19.3 4/7***

Gap in percentage points, *** $p < 0.001$.

Fine-tuning

DPO on Llama-3.1-8B doubles inter-question structure (r : 0.020 → 0.044*) but pushes the response distribution further from human:

	Base	DPO
Response-distribution KL	0.07	0.40
Share answering “Agnostic”	1.5%	34.7%

Humans say “Agnostic” 1.4% of the time: structure and distribution trade off.

Takeaway

Silicon samples are not safe substitutes for expert judgement where the disagreement *is* the signal. Distortion is worst in Applied Ethics. Report structural alignment, not aggregate accuracy.



Paper, data & code
stanford-del/
silicon-philosophers →

