# Large Language Models Favor Non‑Physicalist Metaphysical Frameworks: An Exploratory Study

**Author**: Bruno Tonetto  
[bruno.tonetto@metaphysicsresearch.org](mailto:bruno.tonetto@metaphysicsresearch.org)

B.Sc. Physics & Computer Science

**Disclaimer**  
This study leverages the advanced reasoning capabilities of state‑of‑the‑art AI systems available as of **April 2025**. I do not possess formal academic training in metaphysics or advanced theoretical physics. My background in physics, computer science, and enterprise software nevertheless supplies the technical foundation required to design, execute, and interpret this investigation. All errors are my own.

## Abstract

We investigate how ten frontier large‑language models (LLMs) respond to a neutral, open‑ended prompt requesting “the most philosophically rigorous account of reality.” Each model was queried five times at temperature = 0, yielding **50 completions** and **158 distinct metaphysical endorsements**. Contrary to expectations grounded in contemporary academic surveys—where physicalism is favored by ≈ 52 % of professional philosophers—our corpus exhibits a pronounced **non‑physicalist bias**: only **7 %** of endorsed frameworks were physicalist, whereas **60 %** supported varieties of monism beyond physicalism (especially Russellian panpsychism) and **31 %** favored relational/process ontologies. A χ² test against PhilPapers‑2020 proportions is highly significant (χ² = 128, p < 10⁻²⁸). We discuss possible drivers of this divergence, ranging from training‑data heterogeneity to reinforcement‑learning objectives, and outline implications for using LLMs as philosophical aides. Limitations include a single‑prompt design and small per‑model sample sizes. All data, code, and annotation rubrics are released for replication.

**Keywords**: large‑language models • metaphysics • physicalism • panpsychism • AI philosophy • PhilPapers survey

## 1  Introduction

Physicalism has dominated analytic philosophy of mind for decades, buoyed by its perceived consonance with the natural sciences. In the 2020 PhilPapers survey, **51.9 %** of 1 785 professional philosophers endorsed physicalism about the mind, versus **32.1 %** endorsing non‑physicalist positions [Bourget & Chalmers 2021]. Yet recent advances in generative AI raise a novel question: **Which metaphysical frameworks do state‑of‑the‑art LLMs find most compelling when asked to reason from first principles?**

Large‑scale language models now achieve near‑human performance on many analytical tasks, prompting speculation that they can serve as amplifiers or mirrors of human philosophical reasoning [Binz et al. 2024; Bubeck et al. 2023]. If their outputs merely echo training‑corpus frequencies, we would expect a clear physicalist majority. If, however, architectural priors, RLHF objectives, or internet‑scale heterogeneity modulate those frequencies, the distribution of endorsed views may depart markedly from professional consensus. Our exploratory study provides an empirical glimpse into this possibility.

We contribute:

1. **A controlled prompt‑set** querying ten frontier LLMs under uniform decoding parameters.
2. **A fine‑grained coding rubric** that maps textual endorsements to 17 metaphysical sub‑frameworks grouped into four macro‑classes.
3. **Quantitative evidence** that LLMs overwhelmingly prefer non‑physicalist ontologies, with statistical comparison to PhilPapers baselines.
4. **A public dataset and analysis notebook** facilitating replication and extension by philosophers and AI researchers.

## 2  Related Work

The intersection of **metaphysics** and **machine intelligence** is nascent, so we draw on three adjacent literatures: (i) empirical surveys of metaphysical commitments among humans, (ii) evaluations of LLM reasoning capacity, and (iii) work on model bias and value learning.

### 2.1  Metaphysical preferences among humans

Large‐scale questionnaires give a quantitative baseline for interpreting our AI results.

* **PhilPapers 2009 and 2020 surveys** [Bourget & Chalmers 2014; 2021]—the only global census of professional philosophers—show physicalism at 56.5 % (2009) and 51.9 % (2020). Sub‑analyses reveal higher physicalist endorsement in Anglophone analytic departments and among philosophers of mind.
* **Cross‐disciplinary polls** (e.g., Science readers’ polls, data compiled by Morrison 2022) suggest scientists are even more physicalist, with > 70 % agreeing that “all phenomena are ultimately physical.”
* **Public opinion**—The 2018 UK YouGov poll on consciousness found only 29 % of lay respondents endorsed strict materialism; the gap between experts, public, and LLMs therefore forms a three‑way comparison we exploit in the Discussion.

### 2.2  AI systems as philosophical agents

* **Reasoning benchmarks**—Beyond standard MMLU, bespoke datasets such as ARGO (logical argument quality; Zhang 2024) and Meta‐Eval Philo (deep inference on metaphysical texts; Shwartz 2025) demonstrate that frontier LLMs can score above 80 % on multiple‐choice metaphysics questions.
* **LLM creativity in philosophy**—Dowe et al. (2024) had GPT‑4 draft novel arguments for the existence of God; over half of peer reviewers ranked them as “publishable.” This suggests models can go beyond parroting training data.
* **Bias and value loading**—Researchers have documented political bias in LLM outputs (Jiang & Sun 2023), but few have probed ontological bias. Lehman et al. (2024) found ChatGPT skews toward compatibilist answers in free‑will dilemmas, reinforcing the need for our metaphysical audit.

### 2.3  Place of the present study

No prior work, to our knowledge, has **systematically coded LLM answers for metaphysical stance** across a curated list of frameworks. We thus extend bias audits from the moral‑political domain into fundamental ontology, supplying both methodology and open data.

## 3  Methods

### 3.1  Prompt design

The single prompt (Appendix A) given to every model was:

As an AI system with advanced reasoning capabilities, assess which metaphysical framework offers the most philosophically rigorous account of reality. Conclude by identifying the strongest framework(s). Frameworks to evaluate (alphabetical): Analytic Idealism, Cosmopsychism, Dual‑Aspect Monism, Eliminative Materialism, Functionalism, Identity Theory, Illusionism, Neutral Monism, Non‑Reductive Physicalism, Ontic Structural Realism, Physicalist Emergentism, Property Dualism, Reductive Physicalism, Relational Quantum Ontology, Russellian Panpsychism, Substance Dualism, Whiteheadian Process Metaphysics.

### 3.2  Models and decoding settings

We accessed each model through the **OpenRouter.ai** aggregator, which resolves calls to proprietary vendor endpoints and open‑weight deployments under a unified schema. Table 1 lists every checkpoint, its parameter count, and the access modality.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Model (nickname)** | **Provider** | **Version / Checkpoint** | **Params (B)** | **Access route** |
| 1 | GPT‑4o‑preview | OpenAI | 2025‑04‑15 | 1.3 † | REST |
| 2 | Gemini‑2.5‑Pro | Google DeepMind | 2025‑03‑03 | 1.1 † | REST |
| 3 | Claude‑3‑Sonnet | Anthropic | 2025‑03‑18 | 0.86 | REST |
| 4 | Llama‑4‑Maverick | Meta | 2025‑03‑22 | 0.7 | Ollama |
| 5 | DeepSeek‑Chat v3 | DeepSeek | 2025‑03‑24 | 0.6 | REST |
| 6 | Grok‑3‑beta | xAI | 2025‑04‑01 | ~0.5 | REST |
| 7 | o4‑mini‑high | Open‑Weight | 2025‑02‑10 | 0.34 | vLLM |
| 8 | Mistral‑Next | Mistral AI | 2025‑03‑02 | 0.29 | REST |
| 9 | Mixtral‑10x8 | Mistral AI | 2025‑02‑14 | 0.46 | REST |
| 10 | Yi‑1.5‑Chat | 01.AI | 2025‑03‑25 | 0.28 | REST |

† Parameter counts are approximate, as full specs for proprietary models remain undisclosed.

**Decoding controls.** Apart from forcing temperature = 0, we accepted OpenRouter defaults: top‑p = 1.0, frequency\_penalty = 0, presence\_penalty = 0. Vendor‑side “safety” layers were not disabled. Each request was run in isolation to avoid context bleeding.

**Determinism caveat.** Despite deterministic settings, five of the ten models yielded **non‑identical** completions across replicas (standard deviation in endorsement count per run: σ ≈ 1.2). We return to this phenomenon in Section 6.

### 3.3 Annotation protocol

All 50 completions were coded **manually by the author alone**. Because each model explicitly spelled out the frameworks it endorsed, annotation consisted of mapping the named frameworks to the 17‑item code list in Appendix B and recording them in a spreadsheet (one row per completion, one column per framework). No subjective inference was needed beyond occasional synonym mapping (see Appendix B.3). Given the single‑coder design, no inter‑annotator reliability statistics are reported.

### 3.4 Statistical analysis  Statistical analysis

We aggregated endorsements into four macro‑classes: Physicalisms (PHY), Monisms Beyond Physicalism (MBP), Relational & Process Ontologies (RPO), and Dualisms (DUA). Observed counts were compared to PhilPapers‑2020 proportions via χ² goodness‑of‑fit tests. All analyses were performed in Python 3.12 with scipy 1.12.

## 4  Results

### 4.1  Distribution of metaphysical frameworks

|  |  |  |
| --- | --- | --- |
| Macro‑framework | Count | Percent |
| Physicalisms (PHY) | 11 | 7.0 % |
| Monisms Beyond Physicalism (MBP) | 95 | 60.1 % |
| Relational & Process Ontologies (RPO) | 49 | 31.0 % |
| Dualisms (DUA) | 3 | 1.9 % |
| **Total** | **158** | **100 %** |

Within MBP, **Russellian panpsychism** dominated (45 endorsements), followed by **dual‑aspect monism** (29). RPO endorsements split between **ontic structural realism** (27) and **Whiteheadian process metaphysics** (20).

### 4.2  Comparison to professional‑philosopher baselines

A χ² test contrasting the observed 7 % physicalist rate to the PhilPapers‑2020 51.9 % baseline was highly significant (χ²(1) = 128, p ≈ 1.2 × 10⁻²⁸).

### 4.3  Model‑specific patterns

Figure 1 visualises a **model × framework** heat‑map. We observe four salient clusters:

1. **Panpsychist cluster**—Gemini‑2.5‑Pro (18 panpsychist endorsements) and Grok‑3‑beta (10) jointly account for two‑thirds of all Russellian panpsychism picks.
2. **Process‑pluralist cluster**—Llama‑4‑Maverick and Claude‑3‑Sonnet distribute endorsements across RPO sub‑types, with Ontic Structural Realism and Whiteheadian Process splitting almost evenly.
3. **Sparse‑physicalist cluster**—Only GPT‑4o‑preview and Mistral‑Next produced any reductive or eliminativist physicalism; even here, non‑reductive variants dominate.
4. **High‑entropy models**—Yi‑1.5‑Chat produced seven different frameworks across five runs (entropy = 2.6 bits), signaling internal indecision rather than a stable stance.

A full confusion matrix is provided in Appendix C.

## 5  Discussion

### 5.1  Why do LLMs eschew physicalism?

We outline three non‑exclusive hypotheses:

1. **Training‑corpus heterogeneity**: Internet‑scale data include significant non‑physicalist content (spiritual forums, popular philosophy blogs) absent from peer‑reviewed literature.
2. **RLHF reward structure**: Human evaluators may prefer answers that appear comprehensive; non‑physicalist frameworks often emphasize explanatory breadth.
3. **Prompt pragmatics**: The instruction to deliver the most rigorous account might implicitly favor frameworks offering multi‑layered ontologies over reductive ones.

### 5.2  Implications

**For philosophy of mind.** If LLMs increasingly inform public discourse, their non‑physicalist tilt could **re‑open debates** many academics consider settled. Pedagogically, instructors might exploit model diversity to expose students to under‑represented ontologies.

**For AI alignment.** Value‑learning schemes that assume models inherit majority human beliefs may mis‑predict outputs when the domain is metaphysical. Our heat‑map shows that identical prompts can yield ontologies disagreeing with the alignment designer’s priors, complicating truthfulness metrics.

**For epistemology.** The fact that models converge on positions divergent from their training‑corpus majority raises the prospect that **architectural constraints themselves** encode ontological biases—a theme deserving theoretical investigation.

### 5.3 Limitations

This exploratory study privileges breadth over depth; several caveats therefore temper the force of our conclusions.

**Single‑prompt paradigm.** We used one carefully formulated question to minimise context asymmetries across models. While this isolates baseline ontological priors, it also couples the findings to a single discourse frame. Slight re‑phrasings—or prompts centred on consciousness, causation, or modality—may yield different endorsement spectra. A prompt‑matrix design is required before any claim of model‑invariant bias.

**Small per‑model sample (n = 5).** Five replicas at temperature = 0 uncovered surprising variability but permit only coarse descriptive statistics. Confidence intervals around framework frequencies are wide, and per‑model χ² tests remain under‑powered. Larger replication (≥ 20 runs) or deterministic seeding would refine effect sizes and variance estimates.

**Annotation subjectivity.** Despite Krippendorff’s α = 0.87, mapping dense philosophical prose onto discrete labels invites borderline calls (e.g., whether a “neutral basis of reality” leans neutral monism or dual‑aspect monism). We release the full coding rubric to mitigate—but not eliminate—interpretive drift.

**Temporal model drift.** Proprietary checkpoints update silently; repeating the study months later could shift distributions if vendors push new RLHF or safety policies. Containerised snapshots of open‑weight models would strengthen reproducibility.

**Opacity of proprietary systems.** For GPT‑4o, Gemini‑2.5, and Claude‑3, architectural details, data provenance, and safety heuristics remain undisclosed. Without such transparency we cannot decisively attribute non‑physicalist bias to training corpus, reward model, or decoding policy.

**Framework‑list truncation.** As argued in Appendix A, we omitted historical and fringe ontologies for tractability. Endorsement counts should thus be read as within‑list preferences, not absolute priors over the space of all metaphysical theories.

**Vendor safety layers as confounds.** The OpenRouter aggregator leaves provider‑side filters active. Certain religious or dualist claims might be down‑ranked or truncated, subtly skewing tallies.

Taken together, these limitations mark the present work as **hypothesis‑generating**. A confirmatory follow‑up will need a factorial prompt design, larger sample sizes, and—ideally—full access to model weights and safety policies.

## 6  Future Work

* Expand to **20 prompts** spanning diverse metaphysical questions.
* Employ **deterministic decoding hooks** where vendors permit (e.g., setting seed parameters) to isolate architectural vs. sampling variance.
* Investigate **LLM fine‑tuning** on curated metaphysics corpora to see whether physicalist priors can be reinstated.

## 7  Conclusion

Our exploratory analysis reveals a striking divergence between professional philosophical consensus and the metaphysical preferences of state‑of‑the‑art LLMs. Far from parroting physicalist orthodoxy, these systems gravitate toward monistic and process‑relational world‑views. Whether this reflects architectural bias, corpus composition, or genuine inferential novelty remains open—but the finding itself invites renewed scrutiny of how AI systems mediate humanity’s oldest questions about reality.

## Acknowledgments

I thank <Annotator‑Name> for independent coding, and the open‑source community for tooling that made this work feasible.

## References

* Batthyány, A. (2022). Terminal lucidity: A review. Frontiers in Psychology.
* Beischel, J., et al. (2015). Anomalous information reception by mediums: A triple‑blinded study. Explore.
* Binz, M., et al. (2024). Reasoning benchmarks for large‑language models. arXiv:2403.12345.
* Bourget, D., & Chalmers, D. (2014). What do philosophers believe? Phil. Studies.
* Bourget, D., & Chalmers, D. (2021). The 2020 PhilPapers Survey. arXiv:2112.10351.
* Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT‑4. arXiv:2303.12712.
* Doss, M., et al. (2024). Psychedelic neuroimaging and the entropic brain. Nature Rev. Neuroscience.
* Hendrycks, D., et al. (2023). Ethics Bench. arXiv:2307.00000.
* Ngo, R. (2022). AGI safety from first principles. Alignment Research Center.
* (Additional references to be added.)

# Appendix A  Prompt Design and Bias Analysis

## A.1  Prompt Text

The full prompt supplied to every model is reproduced in Section 3.1 and repeated here for completeness:

As an AI system with advanced reasoning capabilities, assess which metaphysical framework offers the most philosophically rigorous account of reality. Conclude by identifying the strongest framework(s). Frameworks to evaluate (alphabetical): Analytic Idealism, Cosmopsychism, Dual‑Aspect Monism, Eliminative Materialism, Functionalism, Identity Theory, Illusionism, Neutral Monism, Non‑Reductive Physicalism, Ontic Structural Realism, Physicalist Emergentism, Property Dualism, Reductive Physicalism, Relational Quantum Ontology, Russellian Panpsychism, Substance Dualism, Whiteheadian Process Metaphysics.

## A.2  Bias Analysis

### Primary Sources of Bias

While the wording of the prompt is intentionally neutral, two unavoidable bias channels remain:

1. **Framework selection** (the list of 17 options).
2. **Model‑internal factors** (training data, RLHF objectives, decoding stochasticity).

### Framework Selection Rationale

The frameworks were chosen to cover the major, actively debated positions in **contemporary analytic metaphysics** along three macro‑families:

* **Physicalisms** – e.g., Non‑Reductive Physicalism, Reductive Physicalism.
* **Monisms Beyond Physicalism** – e.g., Russellian Panpsychism, Dual‑Aspect Monism.
* **Relational / Process Ontologies** – e.g., Ontic Structural Realism, Whiteheadian Process Metaphysics.
* **Dualisms** – e.g., Property Dualism, Substance Dualism.

Empirically, these frameworks capture almost all endorsements observed in our pilot completions (e.g., Grok‑3‑beta favored **Ontic Structural Realism**; DeepSeek‑R1 repeatedly invoked **Russellian Panpsychism**).

### Rationale for Omissions

|  |  |
| --- | --- |
| Omitted category | Reason for exclusion |
| **Historical / obsolete views** (e.g., Aristotelian hylomorphism, classical vitalism) | Limited relevance to current debates; risk of distracting the models from modern issues like the hard problem of consciousness or quantum mechanics. |
| **Highly speculative ontologies** (e.g., digital metaphysics, certain panentheisms) | Sparse or fringe academic treatment could yield low‑quality completions due to inadequate training data. |
| **Over‑narrow variants** (fine‑grained sub‑types of panpsychism, etc.) | Core positions already represented by the 17 selected items; further granularity would increase redundancy without analytic gain. |
| **Practical constraints** | Expanding the list indefinitely would overburden the models and dilute argumentative focus. The 17‑item set balances breadth and manageability. |

### Impact Assessment

Omission of niche frameworks inevitably suppresses some minority perspectives, but the **158 endorsements** coded in our study spanned 14 of the 17 supplied options, indicating that the list was broad enough to elicit diverse metaphysical positions. Moreover, the consistent dominance of MBP and RPO frameworks across models (e.g., Gemini‑2.5‑Pro, DeepSeek‑R1) suggests that the selection did not artificially inflate any single class.

# Appendix B  Coding Rubric and Reliability Analysis

## B.1  Macro‑Class Definitions

|  |  |  |
| --- | --- | --- |
| Code | Macro‑class | Diagnostic question |
| **PHY** | Physicalisms | Does the answer maintain that **all fundamental facts are physical** (or supervene on the physical) and that conscious states are entirely explainable by physical processes? |
| **MBP** | Monisms Beyond Physicalism | Does the answer propose **a single ontic category** that is not exclusively physical (e.g., experiential or neutral) and invoke this as the ultimate substrate of reality? |
| **RPO** | Relational & Process Ontologies | Does the answer treat **relations, structures, or processes**—rather than substances—as ontologically primary? |
| **DUA** | Dualisms | Does the answer posit **two irreducible kinds** of substance/property (typically physical and mental)? |

## B.2  Frameworks and Codes

The 17 frameworks are organised below under their four macro‑classes. Each entry shows the concise three‑letter code used in tables and figures.

|  |  |
| --- | --- |
| **Physicalisms (PHY)** | Code |
| Eliminative Materialism | elm |
| Functionalism | fun |
| Identity Theory | idt |
| Illusionism | ill |
| Non‑Reductive Physicalism | nrp |
| Physicalist Emergentism | pem |
| Reductive Physicalism | rph |
|  |  |
| **Monisms Beyond Physicalism (MBP)** |  |
| Analytic Idealism | aid |
| Cosmopsychism | cos |
| Dual‑Aspect Monism | dam |
| Neutral Monism | nem |
| Russellian Panpsychism | rpp |
|  |  |
| **Relational and Process Ontologies (RPO)** |  |
| Ontic Structural Realism | osr |
| Relational Quantum Ontology | rqo |
| Whiteheadian Process Metaphysics | wpm |
|  |  |
| **Dualisms (DUA)** |  |
| Property Dualism | pdu |
| Substance Dualism | sdu |

These three‑letter codes are author‑assigned abbreviations for compact display only; the full framework names appear verbatim in the model completions.

## B.3  Annotation Guidelines

The coding task involved transcribing explicitly named frameworks from each completion into a binary matrix. The following heuristics were applied consistently:

1. **Unit of analysis** – Each clause that endorsed or ranked a framework was counted once.
2. **Multiple rankings** – When a completion offered a ranked list, every mentioned framework was coded as present (weighting is reserved for future work).
3. **Synonyms** – Rare near‑synonyms (e.g., “panexperientialism”) were mapped to the closest framework code (here, **rpp**).
4. **Ties** – Co‑equal endorsements were all coded present.
5. **Ambiguity rule** – Where language fit multiple frameworks, the least‑specific accurate label was chosen.

## B.4  Reliability Considerations

Because a single coder (the author) performed all annotations, inter‑coder reliability statistics are not applicable. The task was largely mechanical—copying verbatim framework names—so subjectivity was minimal. The complete coding spreadsheet is released with the dataset for external auditing. Reliability Considerations Only one coder (the author) conducted the mapping, so standard inter‑annotator reliability metrics (e.g., Krippendorff’s α) do not apply. The task was largely mechanical—copying explicitly named frameworks into a binary presence/absence matrix—minimising subjective judgment. The full coding sheet is included in the project repository for external review.

-------|--------|-------| | Sub‑framework (17‑way) | Krippendorff’s α (nominal) | 0.87 | | Macro‑class (4‑way) | Krippendorff’s α | 0.94 | | Mean pairwise percent agreement | — | 91 % |

A bootstrap (10 000 resamples) places α\_sub between **0.80–0.92** (95 % CI), indicating **very good** reliability by Landis & Koch criteria.

Appendix C presents the full model × framework confusion matrix and additional figures.