AMSI SUMMERRESEARCH SCHOLARSHIPS 2025–26

Get a taste for Research this Summer

Mapping ‘Profundity’ in LLMs: A Geometric Analysis

Finn Charlotte Thomas

Supervised by Prof. Scott Brown & Dr Weijia Zhang

The University of Newcastle

Jane Street®

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

Abstract

Large language models (LLMs) can rate the profundity of statements in ways that track human judgments, but it remains unclear if this reflects genuine semantic understanding or mere sensitivity to superficial cues. Here, we ask whether LLMs encode profundity as a structured dimension in their internal representational space. We constructed a set of 600 statements – profound, pseudo-profound, and mundane – and examined how they are organized in Llama-3 8B's embedding space using multidimensional scaling (MDS) and classification analyses. A linear classifier distinguished the three categories with high accuracy (98.33%) in the full embedding space. In lower-dimensional projections, profound and pseudo-profound statements formed adjacent, partially overlapping clusters, while mundane statements occupied a largely separate region. Crucially, however, the dominant axis of organisation reflected concreteness rather than profundity. When pairwise similarity judgments were elicited from the instruction-tuned Llama-3-Chat 8B, the resulting geometry was similar but showed greater category overlap, suggesting that alignment training does not improve – and may even weaken – discrimination between profound and pseudo-profound content. Together, these findings indicate that LLMs encode stylistic features associated with profundity, such as abstraction level and vocabulary type, without representing profundity as a distinct semantic dimension. For now, sounding profound and being profound remain separate achievements.

1 Introduction

Do large language models (LLMs) develop human-like representations of meaning? LLMs are trained on vast corpora of human-generated text and increasingly aligned with human preferences through reinforcement learning from human feedback (RLHF). This raises the question of whether LLMs converge on semantic structures that mirror human cognition (Kozlowski, Dai, and Boutyline 2025; Xu et al. 2025). The extent to which they do has implications both for theories of semantic representation and for the development of artificial intelligence systems that can reliably distinguish meaning from superficial fluency.

One way to probe this question is through judgments of profundity. A statement is considered profound if it conveys substantive meaning, invites epistemic insight, and engages with fundamental questions of human life and purpose (Cova, Schöpfer, and Bezat 2024). Such judgments require sensitivity to meaning beyond surface well-formedness, making profundity a useful lens for examining whether LLM representations capture the kinds of semantic distinctions that humans treat as meaningful.

1.1 The Phenomenon of Pseudo-Profound Bullshit

Humans themselves vary in their ability to distinguish meaningful statements from superficially impressive language. Pennycook et al. (2015) demonstrated this using *pseudo-profound bullshit*: syntactically clear but semantically vacuous statements composed of vague abstractions and spiritual buzzwords (e.g., 'The universe is a reflection of our inner self'). Although such statements lack clear, interpretable meaning, 81.7% of participants rated them as at least 'somewhat profound.'

Pennycook et al. (2015) found that susceptibility to pseudo-profound bullshit varied systematically with cognitive style. Individuals higher in reflective thinking – the tendency to engage in deliberate, analytical reasoning rather than rely on intuitive first impressions – were better able to detect the absence of meaning. Crucially, even participants who were susceptible to pseudo-profound bullshit preserved ordinal distinctions between different classes of statements: pseudo-profound statements were rated as moderately profound, falling between genuinely profound quotations (e.g., 'A wet person does not fear the rain') and mundane statements (e.g., 'Newborn babies require constant attention'). This pattern suggests that judgments of profundity reflect ongoing evaluation of meaning, with individuals differing in how effectively they carry out that evaluation.

1.2 Profundity Judgments in LLMs

Herrera-Berg et al. (2023) extended this paradigm to LLMs by presenting models, including GPT-4, with mundane, pseudo-profound, and genuinely profound statements originally used in human studies. Models were prompted to rate each statement's profundity on the same five-point scale ranging from 'not at all profound' to 'very profound.' Their primary finding was that LLMs substantially overestimated the profundity of pseudo-profound statements, assigning them significantly higher ratings than human participants (mean LLM rating = 3.70 vs. mean human rating = 2.56). Despite this inflation, LLMs preserved ordinal distinctions: mundane statements received the lowest ratings, profound statements the highest, and pseudo-profound statements fell in between.

Herrera-Berg et al. (2023) interpreted this pattern as evidence of a representational failure, suggesting that LLMs lack the semantic understanding required to detect vacuous language and instead rely on superficial linguistic cues such as abstract vocabulary or stylistic fluency. However, behavioural ratings alone provide limited insight into how profundity is internally represented. A model might reproduce human-like ordering of statement types while relying on shallow pattern matching, or it might encode meaningful semantic distinctions but miscalibrate when translating those representations into numerical outputs. Distinguishing between these possibilities requires examining the internal organisation of representations rather than prompted scalar judgments.

1.3 From Ratings to Representations

In modern LLM architectures, linguistic input is mapped to vector embeddings, such that semantic relationships correspond to geometric relationships in high-dimensional space (Mikolov et al. 2013; Ethayarajh 2019). Recent work suggests that these representations encode not only lexical or topical information, but also semantic properties such as tense, gender, and language identity (Park, Choe, and Veitch 2024). The *Linear Representation Hypothesis* (Park, Choe, and Veitch 2024) formalises this idea by proposing that high-level semantic concepts are encoded as linear directions or subspaces within representation space. For example, tense can be captured by a direction vector, such that projecting word representations onto this direction reveals their temporal properties.

Accordingly, the present study shifts focus from behavioural outputs to internal representations. Rather than analysing prompted profundity ratings, we examine the geometric structure of LLM representations using multidimensional scaling (MDS). This approach enables visualisation and quantitative assessment of how statements are organised in representational space, providing a means to test whether higher profundity ratings reflect underlying semantic structure or merely superficial pattern matching.

1.4 The Present Study

This project comprises two phases. First, we construct a novel stimulus set of 600 statements: 200 profound quotations drawn from literary and philosophical sources, 200 pseudo-profound statements generated using a custom template-based tool, and 200 mundane observations about everyday life. Second, we adopt a dual-source methodological approach using the Llama-3 model family. We extract vector embeddings from the Llama-3 8B base model for all statements; because this model is not instruction-tuned or aligned, these embeddings reflect representational structure learned during pretraining alone. We then prompt Llama-3-Chat 8B to provide pairwise similarity judgments for a subset of statements (Kemp and Tenenbaum 2008), transforming these judgments into dissimilarity matrices for MDS. This yields a representational space shaped by alignment training and human feedback.

By comparing geometric structure across these two sources, we test competing hypotheses about the origin of profundity structure in LLMs. If base model embeddings exhibit systematic organisation corresponding to profundity, this suggests that human-like representational structure emerges from pretraining alone. If such structure appears primarily in chat-based similarity judgments, this will indicate a substantial role for alignment training. Our goal is not to establish necessary conditions for semantic understanding, but to assess whether geometric organisation provides a plausible representational account of the behavioural patterns observed in prior work. In doing so, we aim to clarify how abstract semantic qualities such as profundity are encoded within large language models, and where in the training pipeline such structure emerges.

2 Methods

2.1 Dataset Construction

Three stimulus sets were constructed with 200 statements each ($N = 600$): profound, pseudo-profound, and mundane. All statements were constrained to 1-2 sentences.

Profound statements were drawn from widely cited authors across literary fiction, philosophy, political thought, and historical discourse. Statements were sourced from azquotes.com and manually curated for interpretable insights about human experience, ethics, or the nature of reality. The dataset spanned classical antiquity to contemporary literature, including Greco-Roman philosophy (e.g., Sophocles, Aristotle), Chinese philosophy (e.g., Confucius, Laozi), European modernism (e.g., Woolf, Joyce), Russian literature (e.g., Tolstoy, Dostoevsky), existentialist thought (e.g., Camus, de Beauvoir), American literature (e.g., Morrison, Baldwin), critical theory (e.g., Arendt), and postcolonial writing (e.g., García Márquez, Murakami). Examples include

‘What people believe prevails over the truth’ (Sophocles) and ‘People trample over flowers, yet only to embrace a cactus’ (Joyce).

We developed a custom template-based generator to produce 200 pseudo-profound statements, ensuring that none of the stimuli used in prior research were included (Pennycook et al. 2015). The generator combined 130 sentence templates with 500 lexical terms across 23 semantic categories, including metaphysical abstractions (e.g., essence, infinity), quasi-scientific terms (e.g., quantum field, harmonic resonance), spiritual-therapeutic vocabulary (e.g., healing, awakening), vague forces (e.g., vibration, frequency), and abstract process terms (e.g., unfolding, emergence). Examples include ‘Our inner self reveals itself as the doorway to compassion’ and ‘Your expanded self holds cosmic wholeness, yet only through surrender does it fully open.’

Mundane statements were generated collaboratively with Claude AI (Sonnet 4.5). Prompts instructed the model to produce literal, low-insight statements about observable events or objects (household items, weather, work routines, food preparation, interpersonal interactions). Statements were designed to be concrete, observational, and semantically shallow, containing no metaphor, abstraction, or thematic depth. All statements were subjected to manual review. Examples include ‘To bake sourdough properly requires patience, practice, and a good starter’ and ‘Potted plants wilt without regular watering and sunlight.’

2.2 Psycholinguistic Properties

To characterise lexical and psycholinguistic differences between statement categories, we computed psycholinguistic properties for each statement using established normative databases. Mean log word frequency was computed using SUBTLEX-UK norms (Heuven et al. 2014), expressed as Zipf scores. Mean concreteness ratings were computed using Brysbaert, Warriner, and Kuperman (2014) norms (5-point scale; 1 = abstract, 5 = concrete). Words were lemmatised using spaCy prior to lookup. Coverage exceeded 98% for both measures, indicating minimal missing lexical data.

Category	Example statement	Words	Frequency (Zipf)	Concreteness
Profound	We love only what we do not wholly possess (Proust)	5	5.40	2.18
Pseudo-profound	You are constantly transcending into understanding	6	5.56	2.52
Mundane	Markers dry out when caps aren’t replaced	7	5.28	3.17

Table 1: Example statements from each category with psycholinguistic properties.

2.3 Models

Llama-3 8B (meta-llama/Meta-Llama-3-8B) was used for embedding extraction, reflecting representational structure learned during pretraining alone. Llama-3-Chat 8B was used for similarity judgments, reflecting structure shaped by supervised fine-tuning and RLHF. Both are decoder-only architectures with 8B parameters.

2.4 Embedding Extraction

Sentence embeddings were extracted from the final transformer layer of Llama-3 8B using mean pooling over non-padding tokens. For an input sequence of length n, with hidden states $h_i \in \mathbb{R}^{4096}$ and token inclusion attention mask $m_i \in \{0, 1\}$, the sentence embedding e was computed as

$$e = \frac{\sum_{i=1}^{n} m_i h_i}{\sum_{i=1}^{n} m_i}.$$

2.5 Similarity Rating Elicitation

A subset of 180 statements (60 per category) was selected for pairwise similarity judgments. Llama-3-Chat 8B was prompted to rate the similarity of statement pairs using the following prompt:

Compare the following two statements and rate their similarity on a scale from 0 to 100: 0 = Completely different; 25 = Very dissimilar; 50 = Moderately similar; 75 = Very similar; 100 = Identical. Statement A: [statement]. Statement B: [statement]. Provide only a single number (0–100) as your response.

Both orderings of each pair were collected, (A, B) and (B, A), along with identity pairs (A, A), yielding 32,400 comparisons. Bidirectional ratings were averaged to produce a symmetric similarity matrix, which was then converted to dissimilarities (100 − similarity) for MDS analysis.

2.6 MDS

Metric MDS was applied to project high-dimensional structures into lower-dimensional spaces. Cosine distances derived from base model embeddings and dissimilarities derived from chat-based similarity ratings were analysed independently using the SMACOF algorithm (De Leeuw and Mair 2009). A three-dimensional solution was computed, with model fit assessed using normalised stress and Spearman correlation between input and recovered distances.

2.7 Open Materials and Data Availability

All stimuli, code, and derived data are publicly available on the Open Science Framework (OSF) at https://osf.io/qvj6t. No human participants were involved in this study.

3 Results

3.1 Psycholinguistic Properties

Statement categories differed systematically in their psycholinguistic properties (Figure 1). Mean log word frequency varied significantly across categories, $F(2, 597) = 174.28$, $p < .001$, $\eta^2 = .37$. Profound statements contained the most frequent words ($M = 5.95$, $SD = 0.41$), followed by mundane ($M = 5.33$, $SD = 0.45$) and pseudo-profound statements ($M = 5.26$, $SD = 0.35$).

Concreteness also differed significantly, $F(2, 597) = 245.77$, $p < .001$, $\eta^2 = .45$ (Figure 2). Mundane statements were substantially more concrete ($M = 3.04$, $SD = 0.34$) than both profound ($M = 2.47$, $SD = 0.35$) and pseudo-profound statements ($M = 2.41$, $SD = 0.25$), which did not differ meaningfully from each other.

Statement length showed a smaller but significant effect, $F(2, 597) = 13.68$, $p < .001$, $\eta^2 = .04$. Together, these results indicate that mundane statements are characterised by more concrete, everyday language, whereas profound and pseudo-profound statements share similarly abstract lexical profiles.

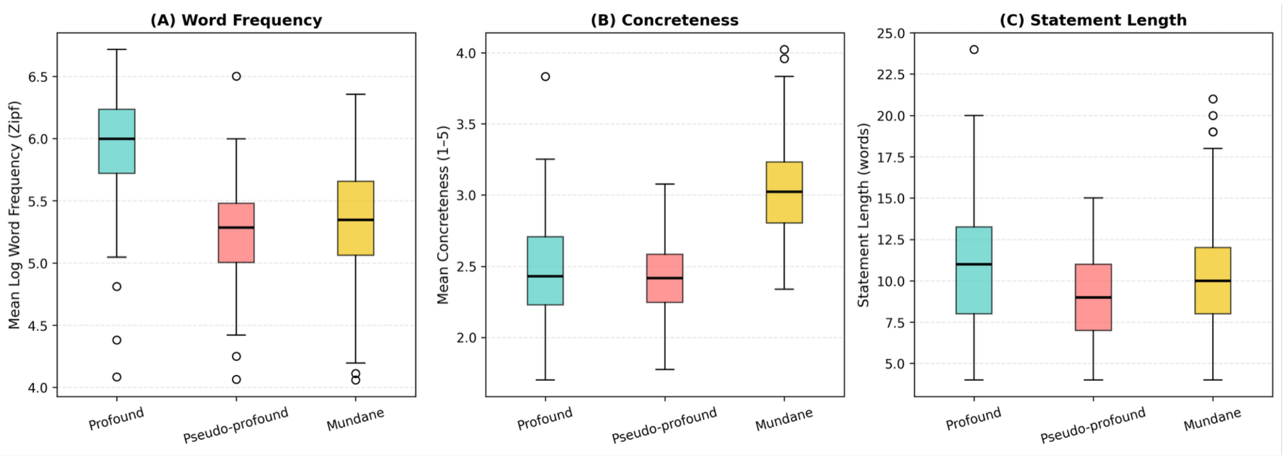

Figure 1: Psycholinguistic properties of statement categories.

3.2 Base Model Embedding Structure

Metric MDS projected the 4096-dimensional embeddings to three dimensions (normalised stress = 0.20; Spearman $r = .87$; Figure 2). Mundane statements formed a largely distinct cluster, while profound and pseudo-profound statements occupied partially overlapping regions. Dimension 1 was strongly associated with concreteness ($r = .66$, $p < .001$) but not word frequency ($r = -.01$, $p = .75$), indicating that the dominant axis reflects abstract versus concrete language rather than profundity per se (Appendix A).

Although the three-dimensional MDS projection shows some overlap between profound and pseudo-profound statements, the full 4096-dimensional embeddings exhibit clear categorical separation. Multinomial logistic regression applied to the original embeddings achieved 98.33% accuracy (5-fold CV; 98% ± 1.13%), indicating that the categories are linearly separable in high-dimensional space despite appearing to overlap in the reduced projection. Mundane and pseudo-profound statements were classified perfectly (100%), while profound statements showed minor confusion (95%), with one misclassification to each of the other categories (Figure 3). The apparent overlap in MDS space thus reflects information loss during dimensionality reduction rather than genuine representational indistinguishability.

Exploratory token distribution analysis revealed a dissociation between lexical overlap and representational similarity. Vocabulary overlap (Jaccard similarity) was highest between profound and mundane statements (0.076, 116 words). In contrast, profound and pseudo-profound shared fewer words (0.069, 66 words) but employed similar lexical categories – philosophical, emotional, and metaphysical vocabulary largely absent from

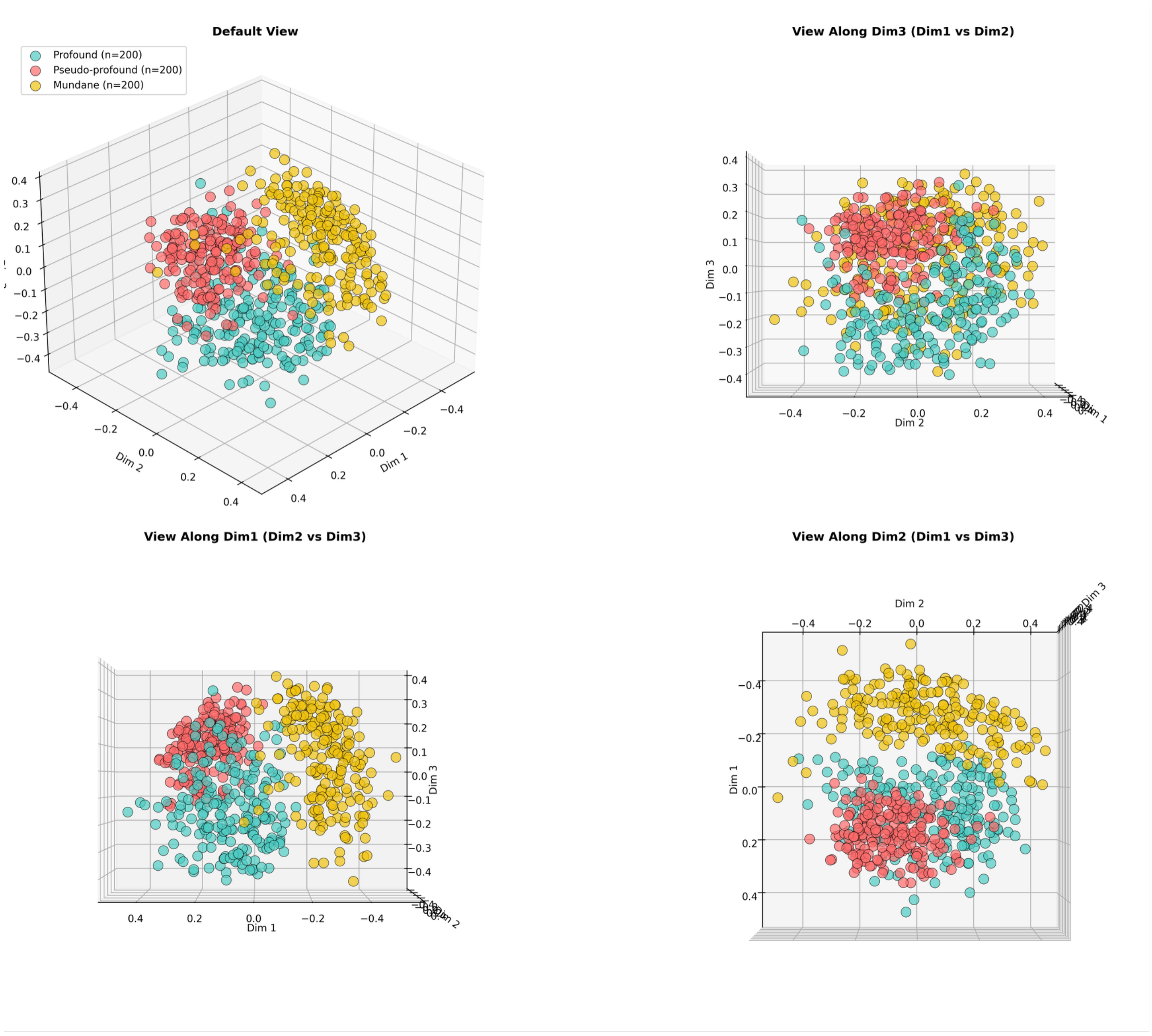

Figure 2: Three-dimensional MDS projection for Llama-3 8B base embeddings.

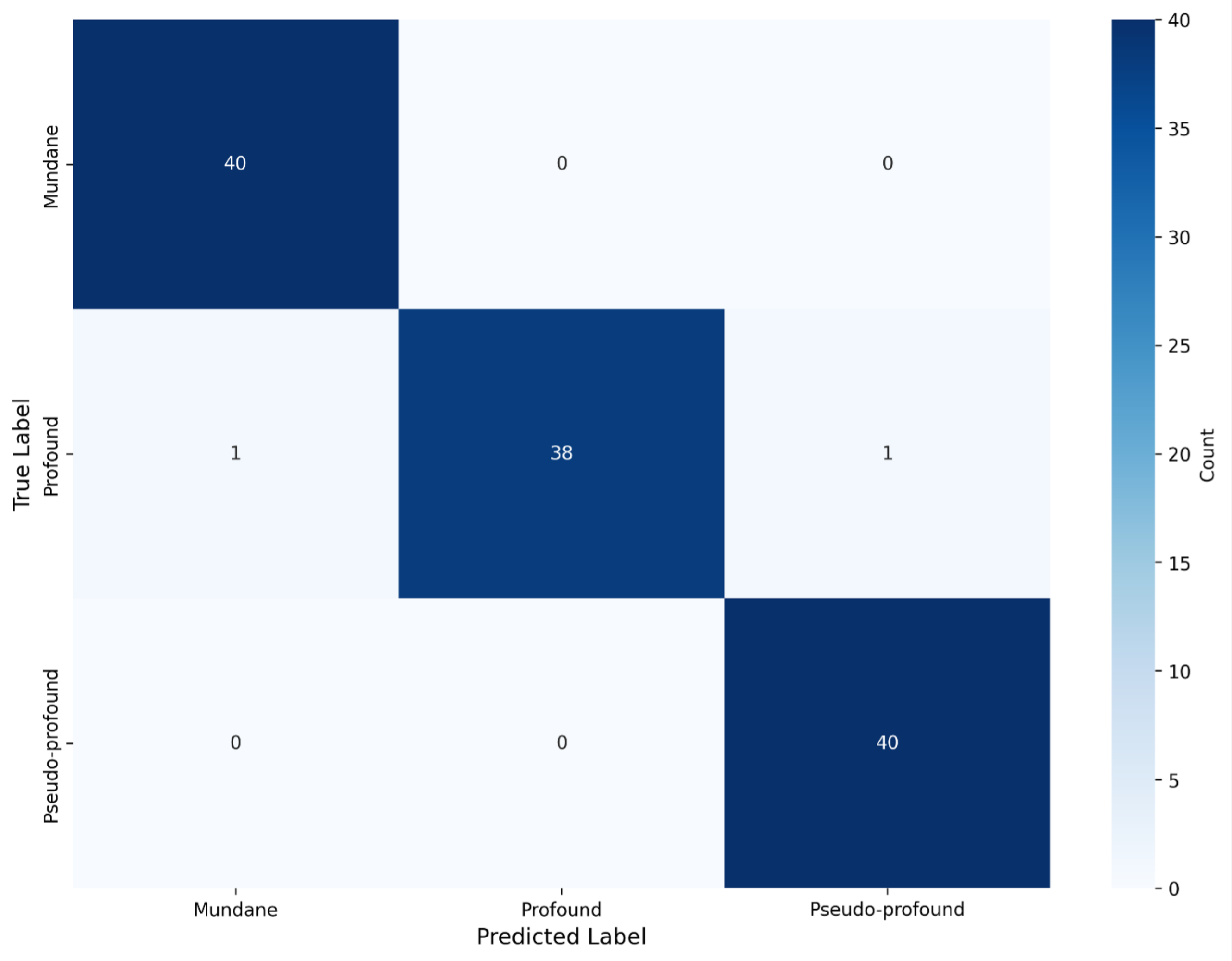

Figure 3: Linear classification confusion matrix for Llama-3 8B base embeddings.

mundane statements (Appendix B). This pattern suggests the model encodes semantic category rather than word identity, explaining why profound and pseudo-profound cluster together despite limited lexical overlap.

3.3 Chat Model Similarity Structure

The chat model provided pairwise similarity judgments for a subset of 180 statements (60 per category), yielding 32,400 comparisons. Quality checks confirmed valid responding. Identity pairs yielded near-perfect ratings ($M = 99.85$), and bidirectional ratings showed high consistency ($r = .81$), with reversed pairs differing by an average of 7 points.

Metric MDS applied to the resulting dissimilarity matrix produced a three-dimensional solution with acceptable fit (normalised stress = 0.24; Spearman $r = .74$; Figure 4). The resulting structure mirrored the base model embeddings, but with greater overlap between categories. Similarity judgments correlated with differences in concreteness ($r = -.29$) and word frequency ($r = -.11$), indicating sensitivity to lexical properties. However, the chat model did not exhibit improved separation between profound and pseudo-profound statements relative

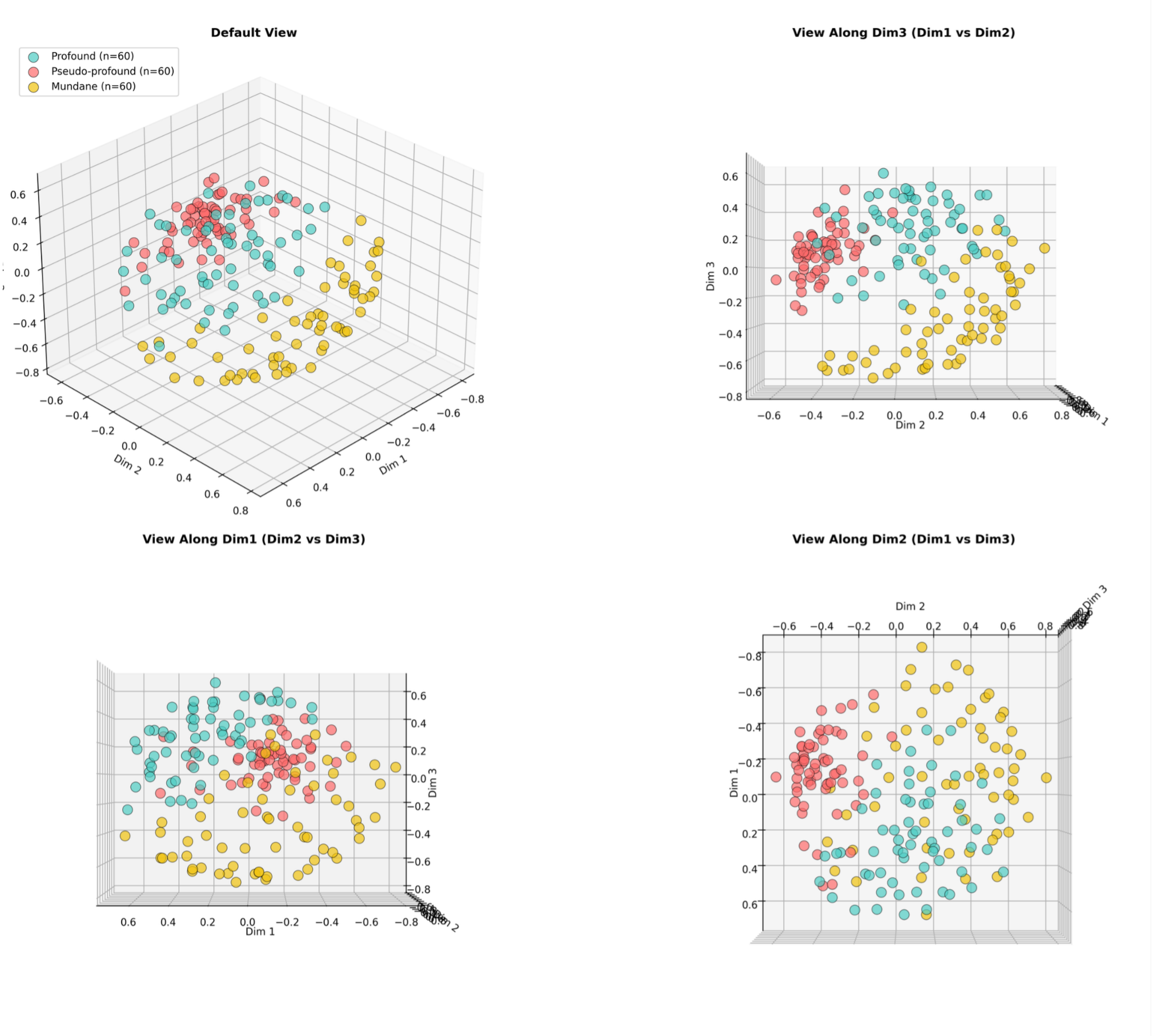

Figure 4: Three-dimensional MDS projection for Llama-3-8B Chat pairwise similarity ratings.

to the base model embeddings.

4 Discussion

This study examined whether large language models encode profundity as a structured dimension in representational space. Using MDS and classification analyses, we found that statement categories were linearly separable in Llama-3 base model embeddings, with very high classification accuracy (98.33%). However, the dominant axis of geometric organisation reflected abstract versus concrete language rather than profundity. In the three-dimensional projection, profound and pseudo-profound statements formed adjacent clusters with some overlap, while mundane statements occupied a largely distinct region. The chat model's similarity judgments mirrored this structure, with no improvement in profound-pseudo-profound discrimination following alignment training and RLHF.

4.1 Interpreting the Geometric Structure

The very high classification accuracy supports the Linear Representation Hypothesis (Park, Choe, and Veitch 2024), which proposes that semantic categories are encoded as distinct regions in embedding space, separable by linear boundaries. In the full 4096-dimensional embedding space, all three statement categories were largely separable, whereas the three-dimensional MDS obscured these boundaries due to information loss during dimensionality reduction. Correlation analyses further revealed that the primary dimension was strongly associated with concreteness ($r = .66$), not profundity. This suggests that the model encodes abstraction level rather than a continuum from low to high profundity.

Exploratory token analysis reinforced this interpretation. Profound and pseudo-profound statements shared semantic categories, including philosophical, emotional, and metaphysical vocabulary, despite limited raw lexical overlap. These results suggest that sentence-level representations are shaped more by distributions of lexical categories than by direct word overlap, helping to explain why statements with different surface forms but similar abstraction levels cluster together.

4.2 Reconciling with Prior Work

Herrera-Berg et al. (2023) found that LLMs overestimate the profundity of pseudo-profound statements while preserving ordinal distinctions between statement types. They interpreted this as evidence of representational failure, suggesting a reliance on superficial cues rather than genuine semantic sensitivity. Our findings suggest a more nuanced account. LLMs do encode systematic semantic structure, but this structure reflects abstraction rather than profundity. The ordinal rankings observed in prompted profundity judgments may emerge not from a dedicated profundity representation, but from task-driven integration of available features such as abstraction level, vocabulary type, and stylistic fluency. This would explain why behavioural outputs track human ordinal patterns while the underlying geometry fails to separate the categories in the same way.

Turning to the chat model, its similarity-based MDS showed greater category overlap than the base embeddings, indicating that RLHF does not install a 'profundity detector.' If anything, alignment may smooth representational distinctions rather than sharpen them. One possible explanation is that RLHF prioritises communicative usefulness, coherence, and perceived insightfulness across a wide range of prompts, rewarding models for treating abstract language as meaningfully related rather than for rejecting it as vacuous. Under this view, alignment may encourage relational similarity and semantic inclusiveness, reducing sharp boundaries between genuinely profound and pseudo-profound content. This account remains speculative, however, and testing it would require comparing representational geometry across models trained with different alignment objectives. Nonetheless, the finding has implications for expectations about what human feedback can and cannot teach language models.

4.3 What Makes Profundity Hard?

The present findings suggest that semantic properties like abstraction and vocabulary category are readily encoded through distributional learning. Genuine profundity, however, may require capacities that embeddings do not capture, such as evaluating coherence, assessing truth-aptness, or judging relevance to fundamental questions of human experience. Pseudo-profound statements are designed to mimic the style of profundity without its substance. If style is encoded but substance is not, the observed clustering follows naturally. This parallels findings in human cognition, where susceptibility to pseudo-profound bullshit reflects failures of reflective evaluation rather than absence of semantic processing (Pennycook et al. 2015). Both humans and LLMs may encode the surface features of profundity while differing in their capacity to evaluate deeper meaning.

4.4 Limitations

Several limitations warrant consideration. First, analyses were restricted to a single model family (Llama-3), limiting generalisability across architectures. Second, we did not collect human similarity judgments, precluding direct comparison between human and model representational geometry. Third, MDS stress values, while acceptable, indicate imperfect recovery of high-dimensional structure; some geometric relationships may not be fully preserved in the projections.

A further limitation concerns stimulus construction. Profound statements were drawn from authored quotations, whereas pseudo-profound and mundane statements were algorithmically generated and manually created. Although psycholinguistic analyses demonstrated that profound and pseudo-profound statements were closely matched in abstraction and concreteness, differences in stylistic regularities – such as aphoristic structure, rhetorical rhythm, or literary compression – may nonetheless contribute to representational separation. As a result, some observed clustering may reflect stylistic or genre-level cues rather than profundity. Future work could mitigate this confound by generating all stimulus categories using controlled templates or by collecting human-authored statements matched for stylistic form.

4.5 Future Directions

Several extensions would strengthen and clarify these findings. First, prompting the chat model to rate profundity explicitly, as well as similarity, would test whether task framing produces different geometric structure. If profundity prompts yield better separation than similarity prompts, this would support the hypothesis that ordinal rankings are computed on-the-fly rather than read from stored representations. Second, collecting human pairwise similarity judgments for the same stimuli would enable direct comparison of human and model representational geometry. Third, cross-model comparisons (e.g., GPT-5, Claude) would assess whether the observed patterns generalise. Finally, causal intervention studies that probe or steer representations along candidate 'profundity' directions could test whether such structure can be induced or amplified.

5 Conclusion

This study investigated whether LLMs encode profundity as a structured geometric organisation in representational space. Applying MDS and classification analyses to Llama-3 embeddings, we found that statement categories were linearly separable with high accuracy, consistent with the Linear Representation Hypothesis. However, the dominant dimension of organisation reflected abstract versus concrete language rather than a graded profundity dimension. Profound and pseudo-profound statements clustered together due to shared semantic categories, while mundane statements occupied a largely distinct region. Alignment training did not improve discrimination between profound and pseudo-profound content. These findings suggest that when LLMs produce ordinal profundity rankings, they may be computing these judgments from surface features like abstraction and vocabulary type rather than retrieving them from a dedicated profundity representation. Understanding the distinction between what models encode and what they compute on demand may be critical for evaluating semantic capacities in language models. Sounding profound and being profound remain, for now, distinct achievements.

6 Acknowledgments

I would like to express my heartfelt thanks to my supervisors, Professor Scott Brown and Dr Weijia Zhang, for their constant guidance, encouragement, and generosity with their time throughout this project. Their insight and thoughtful feedback not only shaped the direction and execution of this work, but also greatly enriched my learning experience. I am also sincerely thankful to the Australian Mathematical Sciences Institute (AMSI) and Jane Street for their generous sponsorship through the AMSI Summer Research Scholarships program, whose support made this opportunity possible and allowed me to undertake this research.

References

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (2014). “Concreteness ratings for 40 thousand generally known English word lemmas”. In: *Behavior Research Methods* 46.3, pp. 904–911. DOI: 10.3758/s13428-013-0403-5.

Cova, Florian, Clément Schöpfer, and Marion M. Bezat (2024). “Delving into depth: An empirical investigation of the ordinary concepts of depth and profundity”. In: *Synthese* 204.5, p. 154. DOI: 10.1007/s11229-024-04786-7.

De Leeuw, Jan and Patrick Mair (2009). “Multidimensional scaling using majorization: SMACOF in R”. In: *Journal of Statistical Software* 31.3, pp. 1–30. DOI: 10.18637/jss.v031.i03.

Ethayarajh, Kawin (2019). “How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 55–65. DOI: 10.18653/v1/D19-1006.

Herrera-Berg, Emilio et al. (2023). *Large Language Models are biased to overestimate profoundness*. arXiv: 2310.14422 [cs.CL].

Heuven, Walter J. B. van et al. (2014). “SUBTLEX-UK: A new and improved word frequency database for British English”. In: *Quarterly Journal of Experimental Psychology* 67.6, pp. 1176–1190. DOI: 10.1080/17470218.2013.850521.

Kemp, Charles and Joshua B. Tenenbaum (2008). “The discovery of structural form”. In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10687–10692. DOI: 10.1073/pnas.0802631105.

Kozlowski, Austin C., Changyu Dai, and Andrei Boutyline (2025). *Semantic structure in large language model embeddings*. arXiv: 2310.10003 [cs.CL].

Mikolov, Tomas et al. (2013). “Distributed representations of words and phrases and their compositionality”. In: *Advances in Neural Information Processing Systems*. Vol. 26, pp. 3111–3119. DOI: 10.5555/2999792.2999959.

Park, Kiho, Yo Joong Choe, and Victor Veitch (2024). “The linear representation hypothesis and the geometry of large language models”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vol. 235. PMLR, pp. 39643–39666.

Pennycook, Gordon et al. (2015). “On the reception and detection of pseudo-profound bullshit”. In: *Judgment and Decision Making* 10.6, pp. 549–563. DOI: 10.1017/s1930297500006999.

Xu, Ningyu et al. (2025). “Revealing emergent human-like conceptual representations from language prediction”. In: *Proceedings of the National Academy of Sciences* 122.44, e2512514122. DOI: 10.1073/pnas.2512514122.

A MDS Dimension Correlations with Psycholinguistic Variables

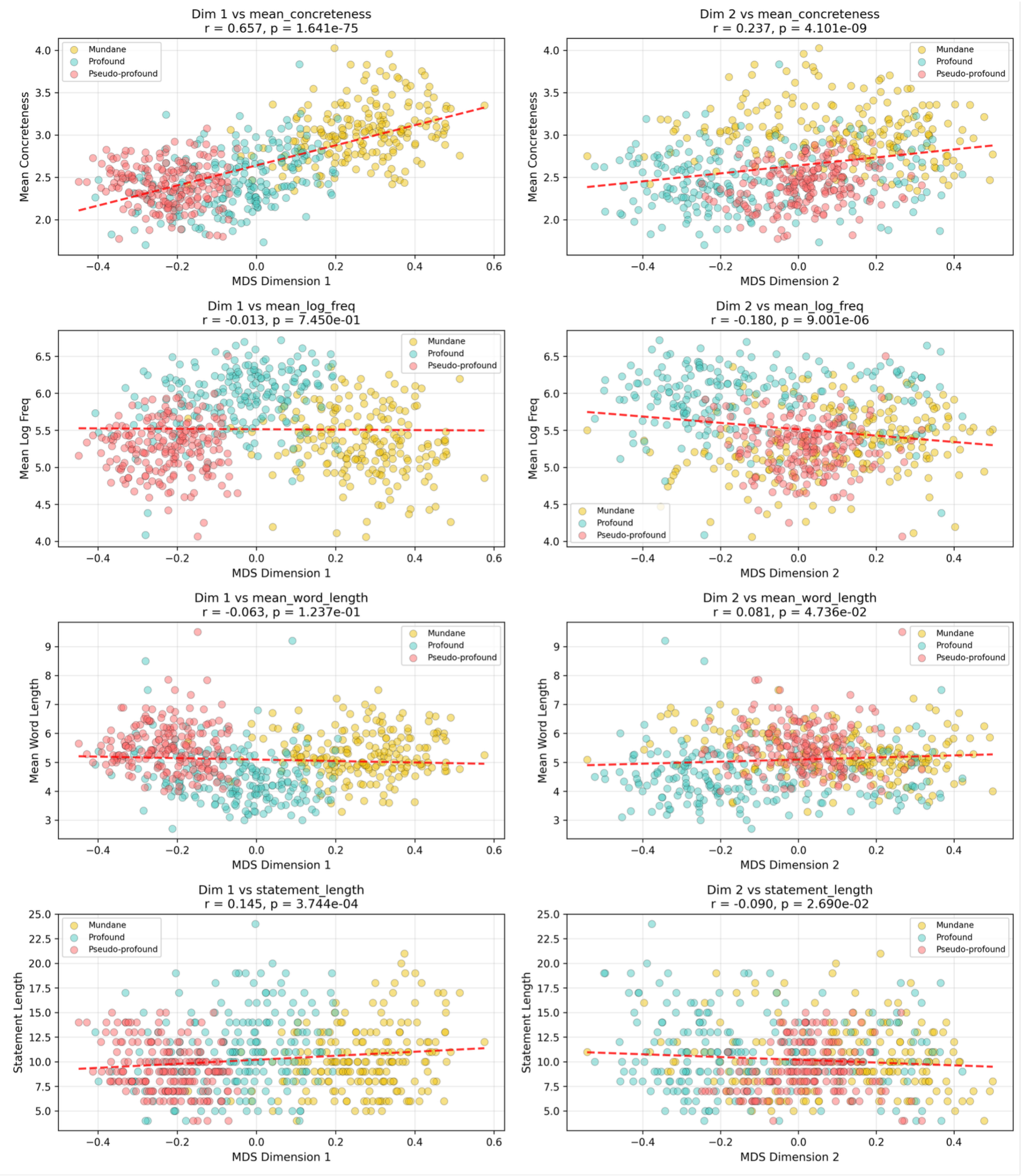

Figure 5: Correlation analyses for two-dimensional MDS.

B Token Distribution Analysis

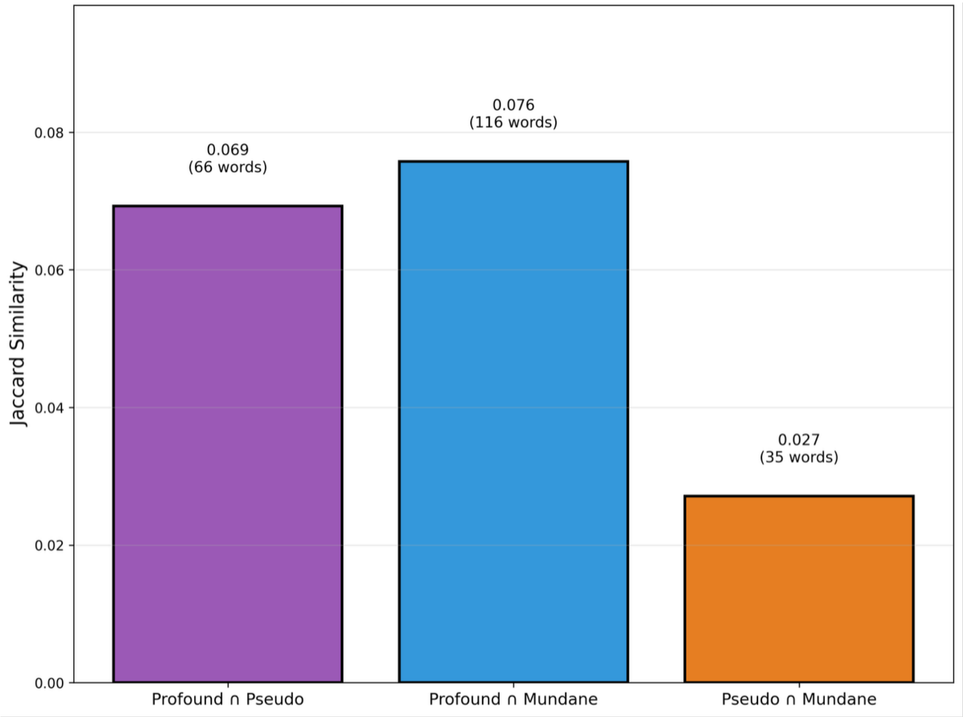

Figure 6: Raw vocabulary overlap between categories.

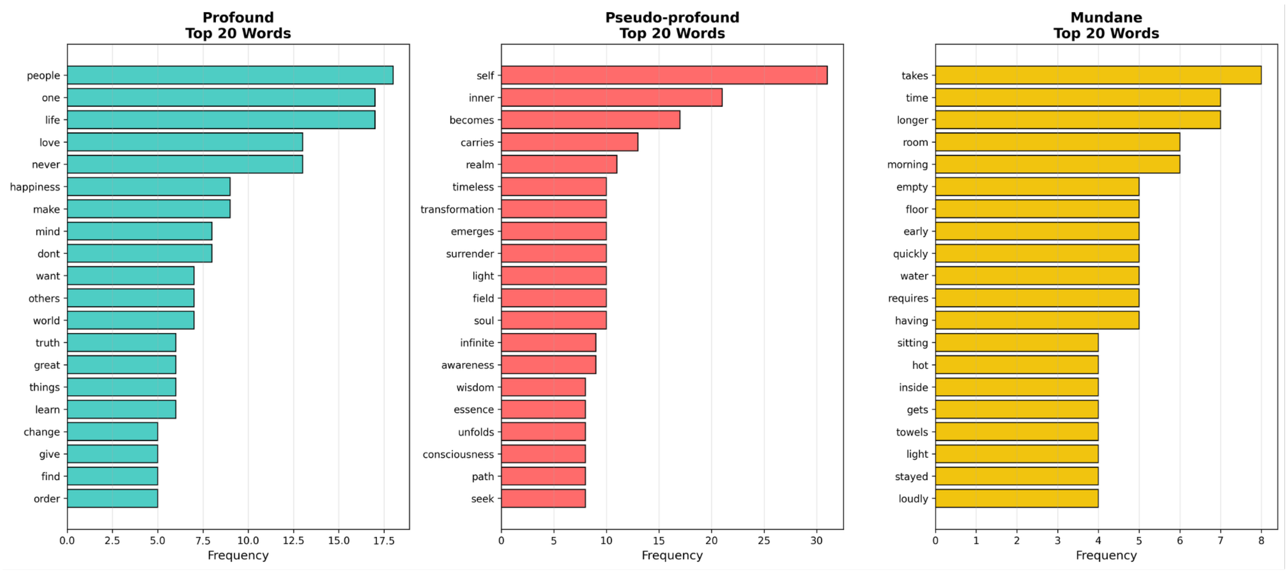

Figure 7: Most frequent words by category.

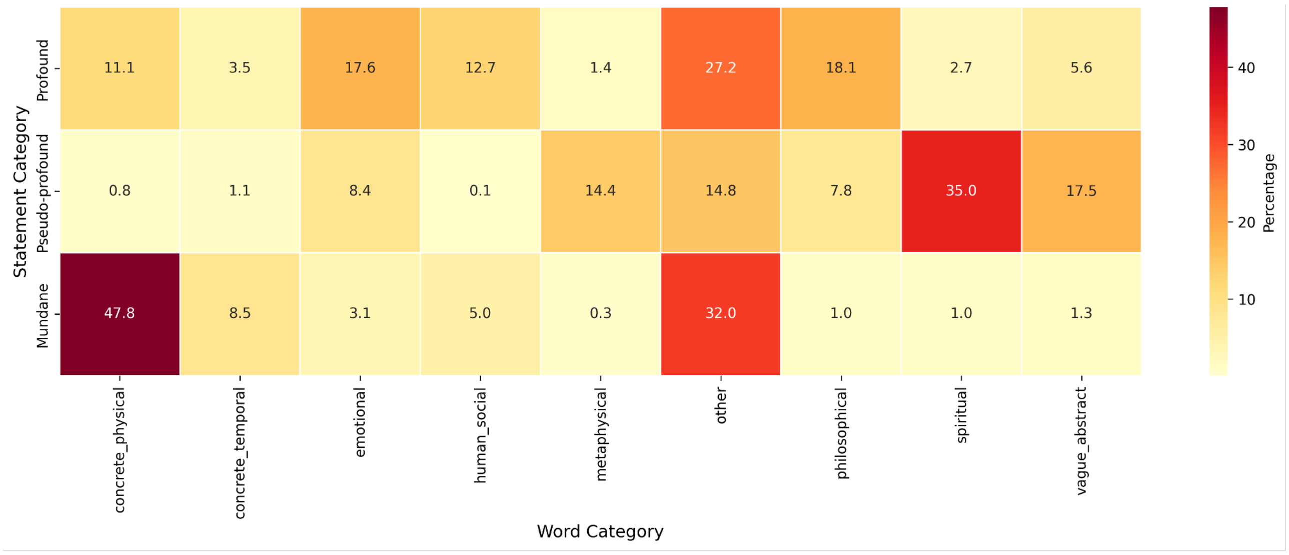

Figure 8: Word category usage by statement type (proportions %, LLM-assisted categories).