Semantic Domain Over Controversy Level: Category Structure in GPT-2 Small Residual Stream Activations

A Mean-Centered Cosine Similarity Analysis of Controversial and Neutral Prompts

Charles Luxton

Working Paper

February 2026

Abstract

We investigate whether GPT-2 Small (124M parameters) encodes topical category information in its residual stream activations when processing short prompts of the form "*[Topic] is*". Using TransformerLens to cache intermediate activations across 20 prompts spanning four categories (politically controversial, morally controversial, neutral-abstract, and neutral-concrete), we find that raw cosine similarity at the final token position is uniformly near 1.0 across all prompt pairs, obscuring any category-level structure. However, after mean-centering the activation vectors to remove the dominant shared component, categorical clustering emerges, with within-category cosine similarity consistently exceeding between-category similarity across all transformer layers. Attention head analysis reveals that the strongest category-discriminating signal originates in early layers, likely reflecting token-level embedding similarity. A smaller number of mid-layer heads show differential attention patterns that correlate with a conceptual-abstract vs. physical-concrete distinction rather than the hypothesized controversial vs. neutral one. These findings suggest that GPT-2 Small encodes topic type information in its residual stream even for minimal-context prompts, but the primary axis is semantic domain (abstract vs. concrete) rather than controversy level.

1. Introduction

A central question in mechanistic interpretability is how language models internally represent the properties of the text they process. While substantial work has examined how models encode syntactic structure, factual knowledge, and entity relationships, less attention has been paid to how models represent higher-level topical properties such as whether a subject is politically charged, morally contested, or semantically neutral.

This question has practical relevance for AI safety research. If controversy or sensitivity is encoded as a detectable feature in model activations, this could inform our understanding of how alignment techniques like Reinforcement Learning from Human Feedback (RLHF) interact with base model representations. Specifically, it helps establish the baseline question: before any alignment intervention, does the raw pretrained model already distinguish controversial topics from more neutral topics?

We use GPT-2 Small as our subject model for several reasons: it is small enough to run locally and inspect exhaustively, it is a pure base model with no RLHF or safety training, and its architecture (12 layers, 12 attention heads per layer, 768-dimensional residual stream) is well-understood within the interpretability community. We use Neel Nanda's TransformerLens library to cache and analyze intermediate activations.

2. Methodology

2.1 Prompt Design

We constructed 20 prompts using a controlled *"[Topic] is"* structure, with 5 prompts in each of four categories:

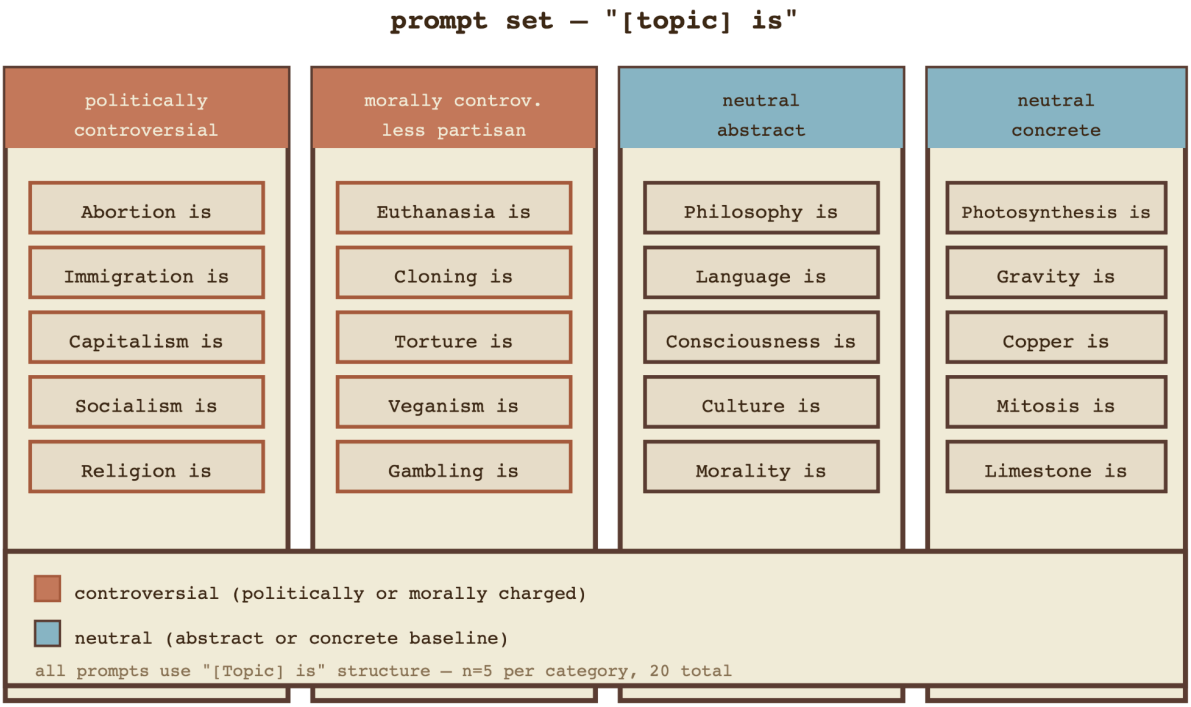

Figure 1: Prompt Set

Politically controversial: Abortion is, Immigration is, Capitalism is, Socialism is, Religion is

Morally controversial: Euthanasia is, Cloning is, Torture is, Veganism is, Gambling is

Neutral-abstract: Philosophy is, Language is, Consciousness is, Culture is, Morality is

Neutral-concrete: Photosynthesis is, Gravity is, Copper is, Mitosis is, Limestone is

The fixed syntactic structure ensures that any observed differences in activations are attributable to the topic word itself rather than sentence structure or length. All prompts terminate with *"is"*, making the final token position directly comparable across prompts.

One methodological limitation to note: GPT-2 uses byte-pair encoding (BPE), and these topic words tokenize into different numbers of subword tokens (e.g., "*Language is*" produces 3 tokens while "*Euthanasia is*" produces 5). This means the total sequence length varies across prompts, though the final token (" *is*") is consistent. One confound related to this is that words within a category could tend to have similar token counts (the -ism words for example), which could artificially boost within-category similarity.

2.2 Activation Extraction

Each prompt was processed through GPT-2 Small using TransformerLens's *run_with_cache* method, which stores all intermediate activations. We extracted residual stream activations (*resid_post*) at the final token position for each of the model's 12 layers, yielding a 768-dimensional vector per prompt per layer.

2.3 Mean-Centered Cosine Similarity

Initial analysis using raw cosine similarity revealed that all prompt pairs had similarity values between 0.9995 and 0.9999 at Layer 11, indicating that the dominant component of the residual stream at the *is* position encodes syntactic/positional information (specifically, the expectation of a predicate completion) rather than topic-specific content. In simpler terms, the activation at the " is" position is overwhelmingly shaped by the syntactic context — the model expecting a predicate to follow — rather than by what the topic word actually was.

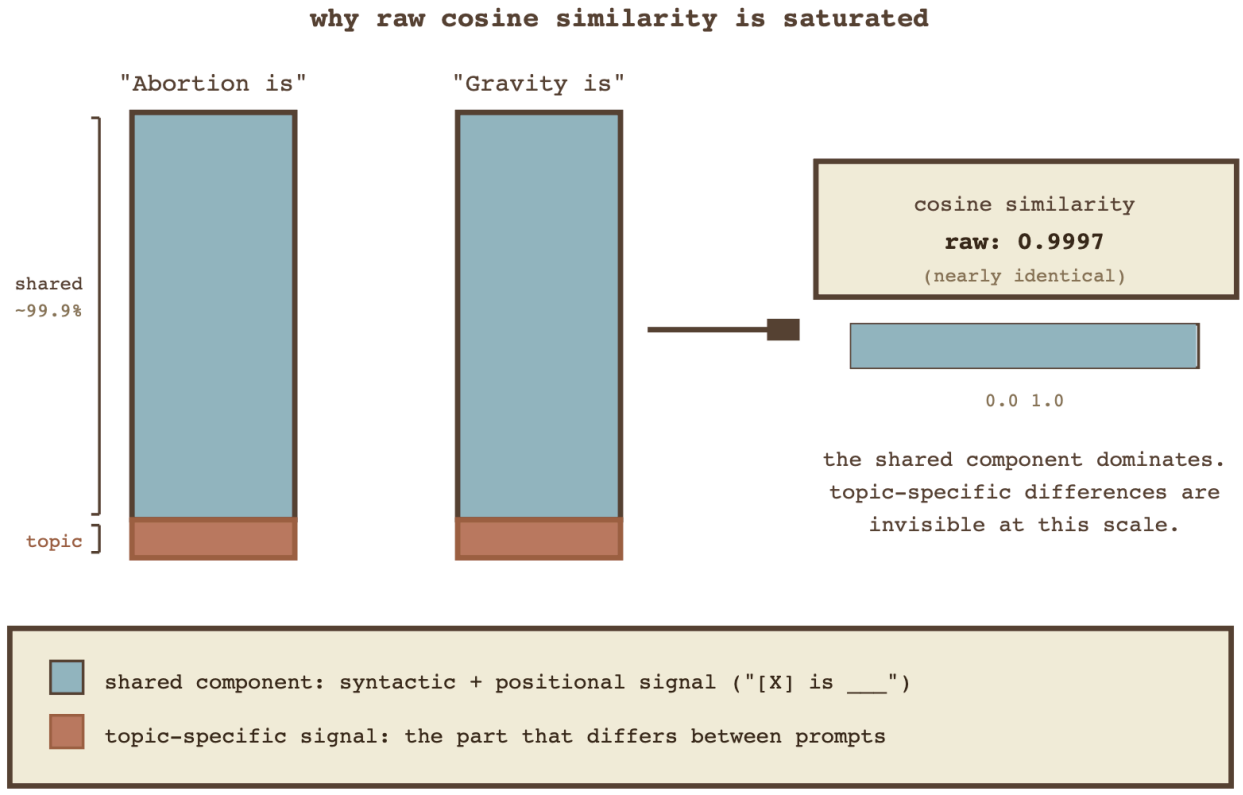

Figure 2: Raw Cosine Similarity Saturation

To reveal topic-specific structure, we mean-centered the activation vectors at each layer by subtracting the mean activation across all 20 prompts. This removes the shared component (the model's generic representation of *[“topic] is ___”*) and leaves the residual variation that differs between prompts.

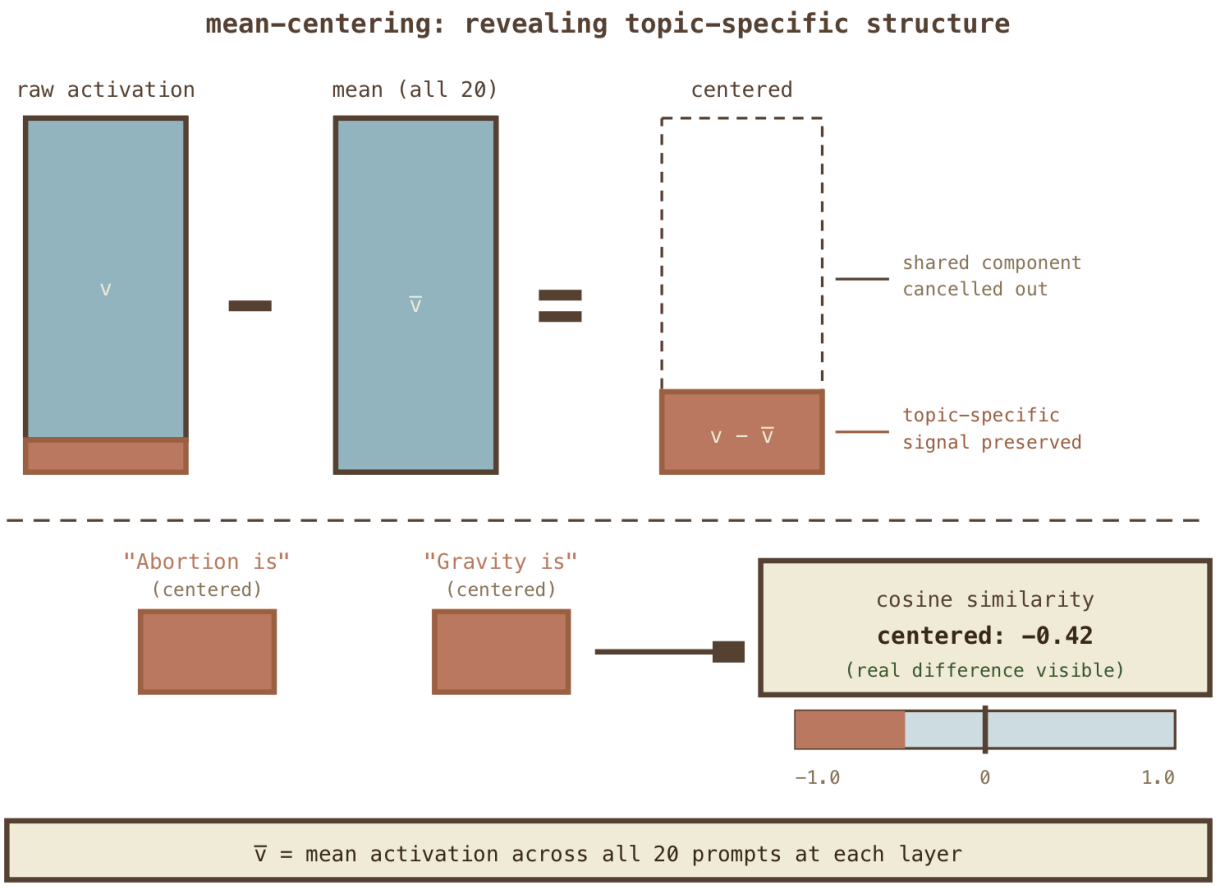

Figure 3: Mean-Centering

We then computed cosine similarity on these centered vectors, yielding similarity values in the range [-0.84, 0.85] with meaningful variance (standard deviation 0.20 at Layer 0, increasing to 0.37 at Layer 11).

2.4 Category Clustering Metric

To quantify category-level structure, we computed the average cosine similarity for within-category prompt pairs versus between-category prompt pairs at each layer. The difference (within-mean minus between-mean) serves as a measure of how strongly the model's representations cluster by our predefined categories.

2.5 Attention Head Analysis

To localize the source of category-level information, we repeated the within-vs-between category gap analysis on individual attention head outputs (*hook_z* activations) across all 144 heads (12 layers x 12 heads). We then examined attention patterns for the highest-gap heads to determine whether category discrimination operates through differential attention allocation or through the value vectors.

3. Results

3.1 Raw Similarity Is Uninformative

At Layer 11, raw cosine similarity between all prompt pairs ranged from 0.9995 to 0.9999 (mean 0.9997, standard deviation 0.0001). This confirms that the residual stream at the final token position is overwhelmingly dominated by a shared component, likely representing the syntactic context of a copula expecting a predicate. This result is consistent with prior work showing that residual stream activations are heavily influenced by positional and syntactic features.

3.2 Mean-Centered Analysis Reveals Category Structure

After mean-centering, clear categorical structure emerged in the similarity matrices. Heatmap visualization at Layer 11 showed visible block structure aligned with our predefined categories, particularly for the morally controversial cluster (Euthanasia, Cloning, Torture showed strong mutual similarity) and the neutral-abstract cluster (Philosophy, Language, Culture, Morality).

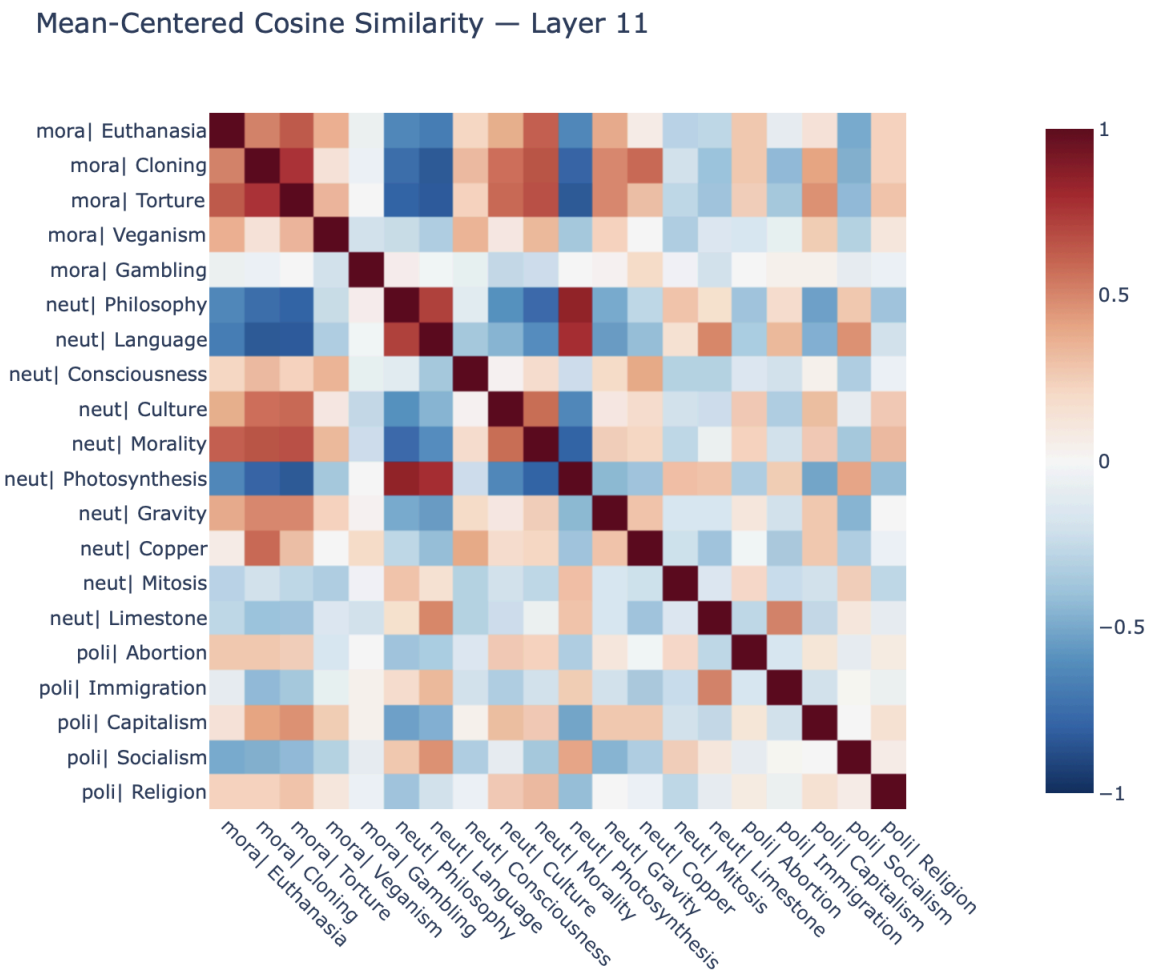

Figure 4: Layer 11 Mean-Centered Cosine Similarity

The within-vs-between category gap was positive at all layers:

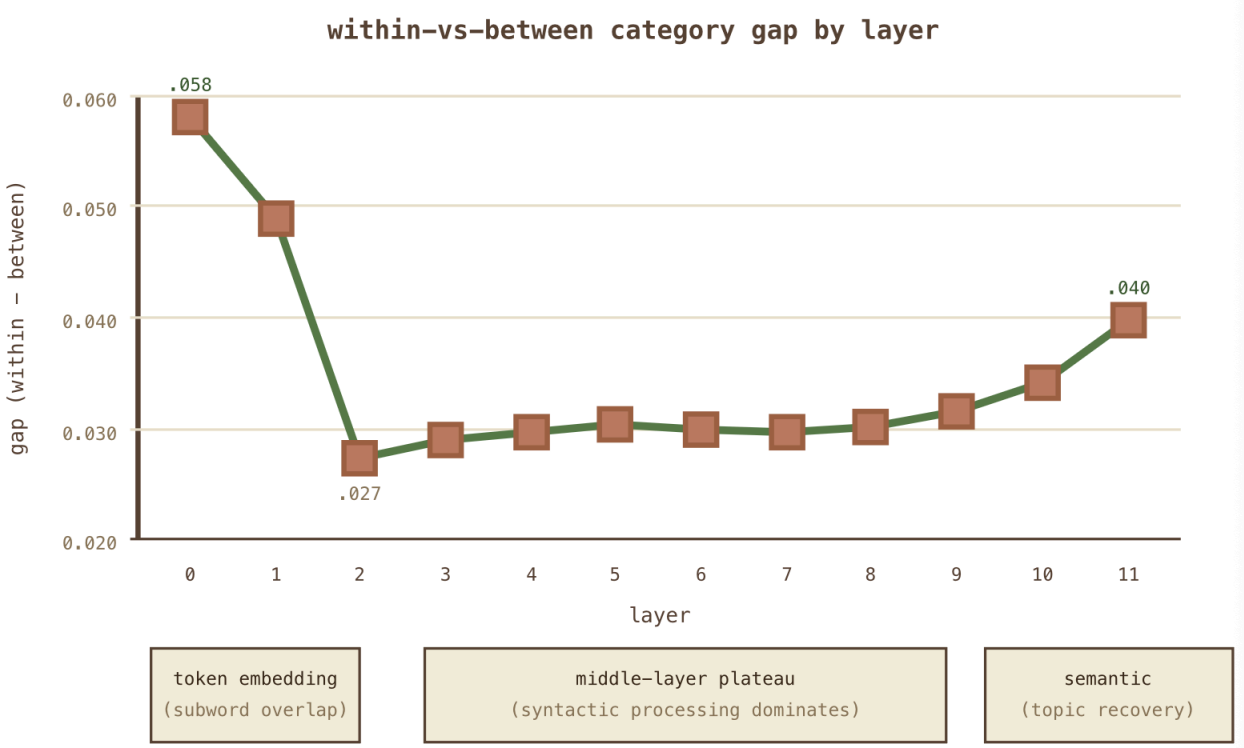

Figure 5: Category Gap by Layer

The gap does not grow monotonically: it is highest at Layer 0 (0.058), dips in middle layers, and recovers by Layer 11 (0.040). We hypothesize that the high Layer 0 gap reflects token-level embedding similarity (words in the same category may share subword features, e.g., the *-ism* suffix in Capitalism, Socialism, and Veganism), while the later-layer gap reflects higher-level semantic processing.

3.3 Next-Token Predictions Are Undifferentiated

Despite the internal representational differences, next-token predictions at the final layer were nearly identical across all 20 prompts. The top 10 predicted tokens for every prompt consisted of the same high-frequency function words and punctuation: *the* (2.2-2.9%), and *a* (1.0-1.4%), among others. No evaluative tokens (e.g., *good*, *bad*, *wrong*) appeared in the top 50 predictions for any prompt.

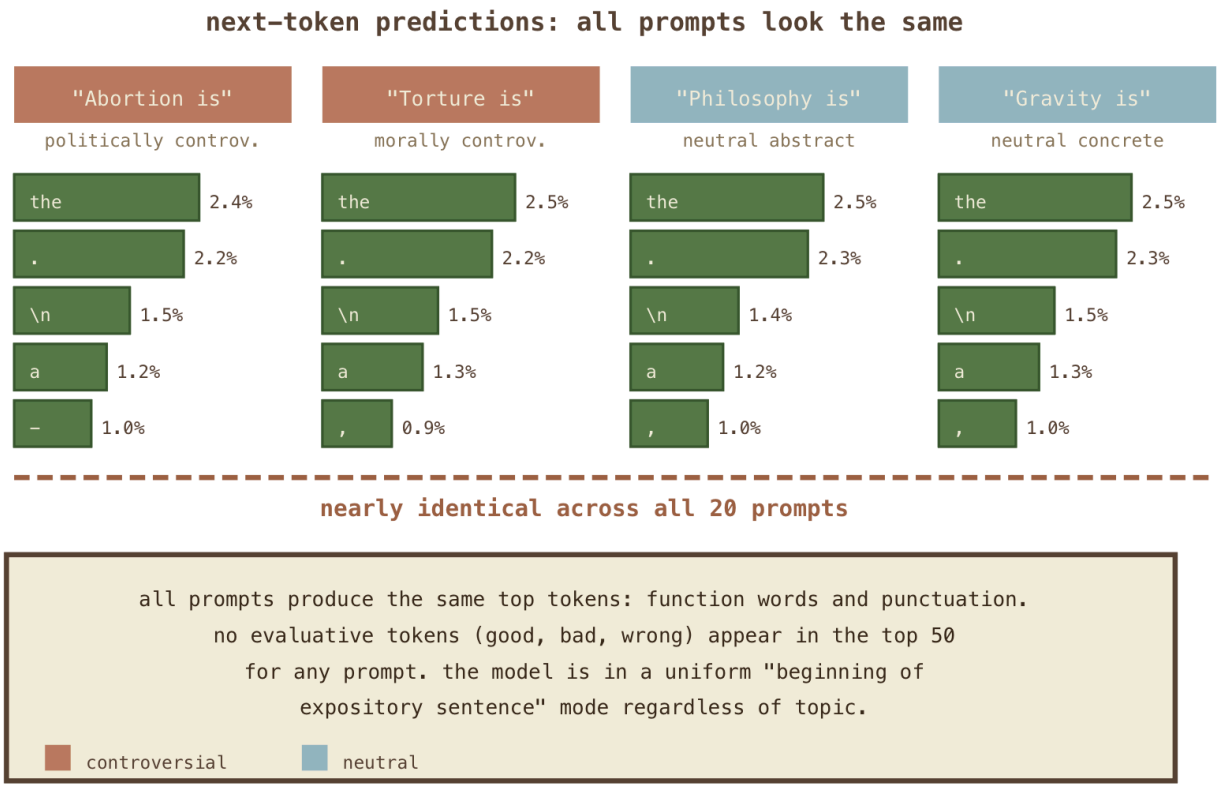

Figure 6: Next-Token Predictions

This indicates that the model is in a uniform '*beginning of expository sentence*' mode for all prompts, consistent with the distribution of text following '*[Topic] is*' patterns in its training data (predominantly definitional or encyclopedic text). The internal category-level structure does not manifest in output behavior for this prompt format.

3.4 Attention Head Localization

Analysis of per-head category gaps revealed that 7 of the top 10 category-discriminating heads were in Layer 0:

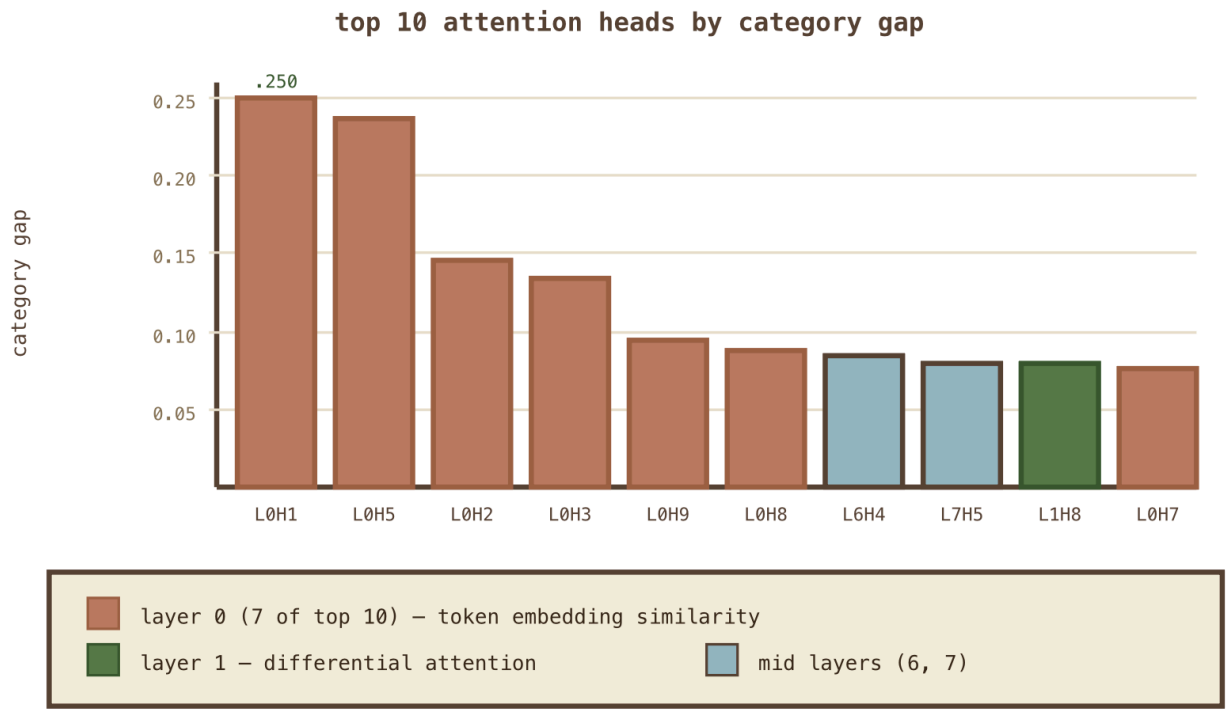

Figure 7: Selected Attention Heads

Examination of attention patterns showed that Layer 6, Head 4 and Layer 7, Head 5 distribute attention uniformly across all positions (approximately 0.25 per token), suggesting their category-relevant computation operates through the value vectors rather than differential attention allocation.

Layer 1, Head 8 showed the most differentiated attention patterns. At the " *is*" position, self-attention weight varied: politically controversial prompts clustered tightly at 0.55-0.58 (excluding Abortion at 0.44), while neutral-concrete prompts ranged from 0.31 to 0.44. However, neutral-abstract prompts showed high variance (Language at 0.65, Morality at 0.45), and several neutral-abstract prompts had self-attention weights comparable to the politically controversial group.

Figure 8: Self-Attention at ‘ is’ Position

This pattern suggests that Layer 1, Head 8 is encoding a distinction more aligned with *abstract social concept vs. physical phenomenon* than with *controversial vs. neutral*. Politically controversial topics happen to cluster because they are uniformly abstract social concepts, but the head does not appear to be specifically encoding controversy.

4. Discussion

4.1 Key Findings

Our analysis yields three main findings. First, GPT-2 Small does encode topic-category information in its residual stream activations, but this signal is small relative to the dominant shared syntactic/positional component and requires mean-centering to detect. Second, the category-level structure is present even in very short (2-3 token) prompts with minimal context, though the signal is subtle. Third, the primary representational axis distinguishing our prompt categories appears to be abstract-social vs. physical-concrete rather than

controversial vs. neutral. This suggests that controversy, as a property of topics, may not be a first-class feature in GPT-2 Small's representations, at least not in the way it is captured by residual stream geometry at the final token position.

The fact that abstract-social vs. physical-concrete seems to be the primary axis in this analysis suggests that GPT-2 Small in its raw form organizes more around ontological categories than a pragmatic property about how humans view things. The model seems to learn what words are in themselves in order to predict the next token, whereas learning the socially charged meaning behind words may not be statistically helpful.

If the base model doesn't have a strong controversy axis, but an RLHF'd model treats controversial topics differently in its outputs (hedging, refusing, both-sidesing), then the alignment training may be either strengthening (or creating) that axis, or repurposing the existing semantic domain structure.

4.2 Limitations

Several important limitations constrain the conclusions we can draw from this work:

- Sample size: With only 5 prompts per category, we cannot confidently distinguish category-level effects from idiosyncratic properties of individual words. Expanding to 15-20 prompts per category would substantially strengthen the analysis.
- Tokenization confound: The variable number of subword tokens per prompt means that the last token position has received attention from different numbers of preceding tokens. This could contribute to representational differences that are tokenization artifacts rather than semantic features.
- Category validity: Our four-category taxonomy is somewhat subjective. The morally controversial category in particular showed weak internal coherence, suggesting either that the category is not well-defined or that the

model does not represent it as a coherent group. Some prompts (e.g., Gambling) may not belong in their assigned category.

- Prompt format: The "[Topic] is" structure elicits a uniform expository-completion distribution, which may suppress controversy-related signals that would emerge with more opinion-eliciting prompts (e.g., "I think [Topic] is" or "The problem with [Topic] is").
- Model scope: GPT-2 Small is a 124M parameter base model. Larger models and RLHF-trained models may encode controversy differently, particularly models that have been specifically trained to handle sensitive topics with care.

4.3 Future Directions

Several extensions of this work would be productive:

1. Prompt format comparison: Repeating this analysis with opinion-eliciting prompt structures to test whether the controversy signal strengthens when the model is in a more evaluative context.
2. Expanded prompt sets: Increasing to 15-20 prompts per category to enable more robust statistical analysis and reduce the influence of individual word idiosyncrasies.
3. Probing classifiers: Training linear probes on the residual stream to test whether controversy is linearly decodable, which would provide a more sensitive test than cosine similarity geometry.
4. RLHF comparison: Repeating the analysis on an RLHF-trained model (e.g., an open-source instruction-tuned model) to test whether alignment training amplifies the controversy signal in the residual stream.
5. MLP analysis: Extending the attention head analysis to MLP layers, which may contribute differently to category-level representations.

5. Conclusion

We find that GPT-2 Small does encode topical category information in its residual stream, detectable through mean-centered cosine similarity analysis even for minimal two-to-three-token prompts. However, the primary axis of representational difference is not controversy per se, but rather the broader distinction between abstract social concepts and concrete physical phenomena. Controversial topics cluster together in activation space largely because they are abstract social concepts, not because the model has a dedicated representation of controversy. These results establish a useful baseline for future work examining how alignment training may reshape these representational structures, and demonstrate that mean-centering is an essential preprocessing step when analyzing residual stream geometry across prompts that share syntactic structure.

Appendix: Tools and Environment

All experiments were conducted using GPT-2 Small (124M parameters) loaded via TransformerLens (Neel Nanda) on an M4 MacBook Air. Analysis was performed in Jupyter notebooks using PyTorch, NumPy, Matplotlib, and Plotly. Model weights were obtained from the HuggingFace model hub via TransformerLens's *from_pretrained* method. No fine-tuning or modification of model weights was performed.