Mechanistic Interpretability of Semantic Abstraction in Biomedical Text

Abstract

We look into whether biomedical language models create register-invariant semantic representations of sentences—a cognitive ability that allows consistent and reliable clinical communication across different language styles. Using aligned sentence pairs (technical vs. plain language abstracts that mean the same thing), we analyze how BioBERT, SciBERT, Clinical-T5, and BioGPT react to varying registers through similarity measures, trajectory visualization, and activation patching. The results show that models converge to shared semantic states in mid-to-late layers, revealing the internal processes by which these models keep meaning across stylistic variation.

1 Introduction & Motivation

Biomedical communication requires translating technical content into plain language without altering meaning, yet how biomedical LLMs represent such semantic abstraction remains unclear, with misrepresentation risking distortion and reduced clinical trust. Prior work shows transformer layers progress from surface features to abstract semantics, but this shift has not been examined in biomedical models or under stylistic variation. We ask: How do biomedical LLMs represent semantically equivalent sentences, and which components preserve meaning across registers? Using aligned pairs from the PLABA dataset (Attal et al., 2023), we analyze BioBERT (Lee et al., 2020), Clinical-T5 (Lu et al., 2022), SciBERT (Beltagy et al., 2019) and BioGPT (Luo et al., 2022). Through representational similarity, attention comparison, and causal probing, we locate depths and components where technical and plain-language inputs converge, offering a mechanistic view of semantic preservation in biomedical NLP.

2 Approach

2.1 Models & Dataset

We analyze four representative biomedical LLM architectures:

- **BioBERT** (encoder-based)
- **SciBERT** (encoder-based)
- **Clinical-T5** (encoder–decoder)
- **BioGPT** (decoder-only)

The **PLABA dataset** provides aligned technical and plain-language biomedical sentences, serving as a natural experiment in semantic stability under register change.

2.2 Layerwise Representation Analysis

For each model, we extract hidden states at every transformer layer and compute: Cosine similarity (Manning et al., 2008), Euclidean Distance, Centered Kernal Alignment (CKA) (Kornblith et al., 2019), Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), and Canonical Correlation– based metrics (SVCCA, PWCCA) (Raghu et al., 2017; Morcos et al., 2018).

2.3 Trajectory & Attention Analysis

We visualize representational trajectories with PCA (Jolliffe, 2011) and t-SNE (van der Maaten and Hinton, 2008), defining them as the layerwise evolution of sentence representations. By comparing the shapes and endpoints of paired trajectories, we assess whether models follow similar abstraction paths across registers. Self-attention maps (Vig and Belinkov, 2019) are analyzed with overlap measures, with semantically analogous tokens aligned via embedding-based cosine mapping to enable direct comparison of attention on technical and plain-language terms.

2.4 Causal Component Analysis Through Activation Patching

We implement a three-stage activation patching pipeline: (1) token alignment via embedding similarity and Hungarian matching, (2) donor bank construction from technical-sentence activations, and (3) patched forward passes with architecture-specific metrics. Evaluation is architecture-specific: cosine similarity for encoder-only models, seq2seq loss for encoder-decoder, and causal LM loss for decoder-only. Donor banks store full-layer and attention-head activations, selectively patched into plain-language streams to reveal components essential for semantic preservation. Loss functions mirror training objectives (MLM, seq2seq, causal LM). To address length mismatches, tokens are aligned by cosine similarity, and activations are patched across heads, MLPs, blocks, and cross-layer combinations to uncover distributed semantic-preservation patterns.

2.5 Experimental Controls

We implement comprehensive validation through three control categories: random activation patching with equivalent dimensionality vectors, semantic control pairs from unaligned biomedical domains, and architectural controls using scrambled connections. Clinical relevance is validated through correlation with medical professional ratings on semantic equivalence.

3 Expected Outcomes

We predict CKA similarity above 0.85 in layers 8-12 for BioBERT, 6-10 for Clinical-T5 encoder, and 12-18 for BioGPT, as prior work shows that transformer models converge to shared semantic representations in mid-to-late layers (Kumar et al., 2024). Trajectory visualization should show converging paths in later layers with technical-plain pairs clustering together. We also expect MLP components to show stronger patching effects than attention heads.

4 Results

4.1 Representation Similarity Across Registers

Across models, similarity analyses showed technical and plain inputs converge in middle-to-late layers. Cosine, RSA, and CKA curves rose to a plateau, indicating early layers capture surface features while deeper layers encode shared semantics.

- **BioBERT and SciBERT** (encoder-only): stable by layers 8–12 (CKA > 0.85), Average Cohen's d per-layer per-neuron of around 0.16.
- **Clinical-T5**: encoder convergence at layers 6–10, Average Cohen's d per-layer per-neuron of around 0.13.
- **BioGPT** (decoder-only): stabilization at layers 14–18. Average Cohen's d per-layer per-neuron of around 0.22; the notably higher value means that this model treats the register-varying sentences more differently.

These results indicate that biomedical LLMs progressively eliminate stylistic variation and converge to register-invariant semantic states in middle-to-late layers.

4.2 Trajectory and Attention Analysis

Trajectory visualizations showed that technical and plain-language pairs diverged in shallow layers but converged later. This indicates that lexical differences are abstracted into equivalent representations. Attention analysis (Figures 1–8) revealed mid-layer heads consistently attending to biomedical entities across registers, with stronger alignment than in shallow or final layers. These results support the hypothesis that convergence emerges in middle-to-late layers.

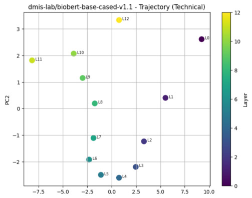

Figure 1: BioBERT trajectory by layer for technical sentence

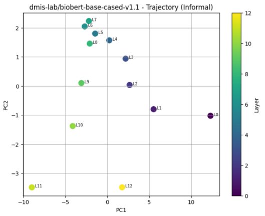

Figure 2: BioBERT trajectory by layer for informal sentence

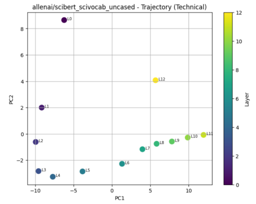

Figure 3: SciBERT trajectory by layer for technical sentence

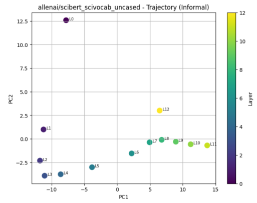

Figure 4: SciBERT trajectory by layer for informal sentence

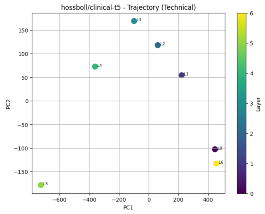

Figure 5: T5 trajectory by layer for technical sentence

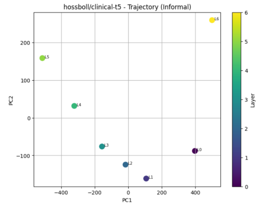

Figure 6: T5 trajectory by layer for informal sentence

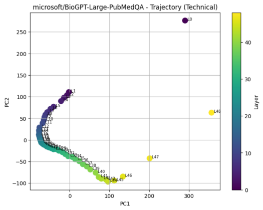

Figure 7: BioGPT trajectory by layer for technical sentence

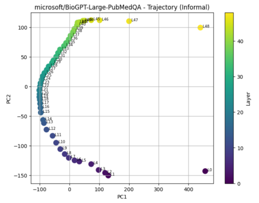

Figure 8: BioGPT trajectory by layer for informal sentence

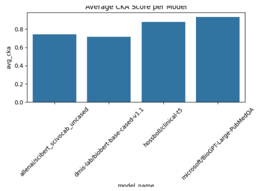

Figure 9: Average CKA score per model

4.3 Causal Component Analysis via Activation Patching

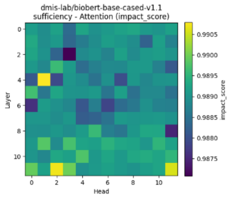

Figure 10: BioBERT Sufficiency Heatmap

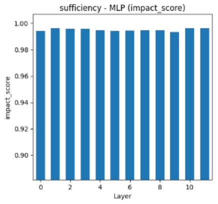

Figure 11: BioBERT Sufficiency Barplot

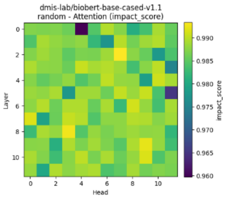

Figure 12: BioBERT Random Heatmap

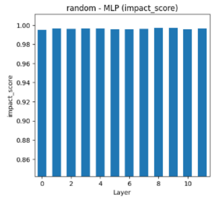

Figure 13: BioBERT Random Barplot

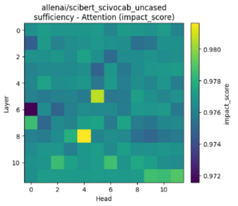

Figure 14: SciBERT Sufficiency Heatmap

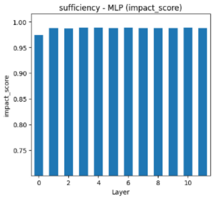

Figure 15: SciBERT Sufficiency Barplot

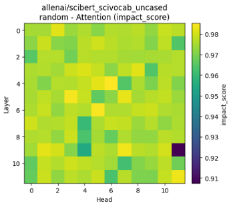

Figure 16: SciBERT Random Heatmap

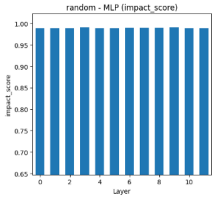

Figure 17: SciBERT Random Barplot

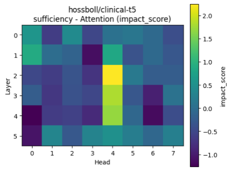

Figure 18: T5 Sufficiency Heatmap

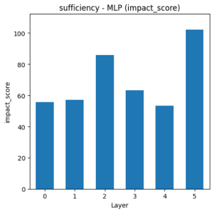

Figure 19: T5 Sufficiency Barplot

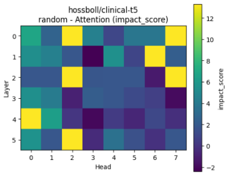

Figure 20: T5 Random Heatmap

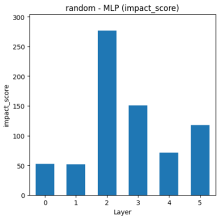

Figure 21: T5 Random Barplot

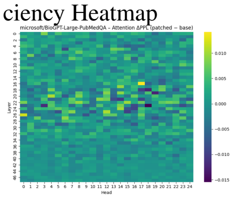

Figure 22: BioGPT Sufficiency Heatmap

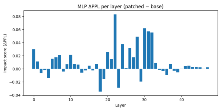

Figure 23: BioGPT Sufficiency Barplot

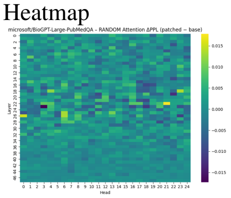

Figure 24: BioGPT Random Heatmap

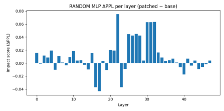

Figure 25: BioGPT Random Barplot

Activation patching experiments, performed on the testing split (Figures 10-25), identified components that causally preserve semantic equivalence under register change.

- **MLPs vs. Attention:** Heatmaps (Figures 10, 11, 14, 15, 18, 19, 22, 25) show MLPs exert stronger causal effects—especially mid-to-late layers—while attention heads are weaker but still above random.
- **Random baselines:** Scrambled activations (Figures 12, 13, 16, 17, 20, 21, 24, 25) yield near-zero effects, confirming sufficiency scores capture genuine register-invariant features.
- **Cross-model trends:**
  - **BioBERT/SciBERT:** Mid-layer MLPs (L8–12) dominated register-invariant encoding.
  - **Clinical-T5:** Distributed sufficiency across encoder layers, without a single sharp peak.
  - **BioGPT:** Register-sensitive MLPs concentrated in deeper layers (L14–18).

    These findings converge with similarity analysis, supporting that semantic preservation emerges in mid-to-late layers and is disproportionately mediated by MLP blocks.

4.4 Summary of Findings

Taken together, similarity analysis, trajectory alignment, and activation patching provide convergent evidence that biomedical LLMs develop register-invariant semantic representations. Encoder-only models converge earlier, decoder-only models later, and MLP components play a stronger causal role than attention heads. These insights explain how biomedical LLMs handle stylistic variation, providing a foundation for interpretable and trustworthy clinical communication systems.

5 Conclusion

This framework identifies the model components that preserve meaning across styles in biomedical language models. These mechanisms could be leveraged in medical question-answering systems through targeted fine-tuning while tracking efficiency and accuracy to semantic representation. Future work will extend these findings to patient-clinician interactions and multilingual biomedical settings, enabling us to further characterize the internal algorithms that facilitate robust, register-independent semantic processing.

References

Attal , K., Ondov , B., & Demner-Fushman , D. (2023) A dataset for plain language adaptation of biomedical abstracts. *Scientific Data* **10**(1):8.

Beltagy , I., Lo , K., & Cohan , A. (2019) Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* pages 3615–3620.

Jolliffe , I. Principal component analysis. *International Encyclopedia of Statistical Science*. Springer, 2011.

Kornblith , S., Norouzi , M., Lee , H., & Hinton , G. (2019) Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning* pages 3519–3529.

Kriegeskorte , N., Mur , M., & Bandettini , P. A. (2008) Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**:4.

Kumar , S., Sumers , T. R., Yamakoshi , T., & al. (2024) Shared functional specialization in transformer-based language models and the human brain. *Nature Communications* **15**.

Lee , J., Yoon , W., Kim , S., Kim , D., Kim , S., So , C. H., & Kang , J. (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4): 1234–1240.

Lu , Q., Dou , D., & Nguyen , T. H. (2022) Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022* pages 5094–5106.

Luo , R., Sun , L., Xia , Y., Qin , T., Zhang , S., Poon , H., & Liu , T.-Y. (2022) Biogpt: Generative pre-trained transformer for biomedical text generation and mining. In *Briefings in Bioinformatics*

Manning , C. D., Raghavan , P., & Schütze , H. (2008) *Introduction to information retrieval*. *Introduction to information retrieval*: Cambridge university press.

Morcos , A. S., Raghu , M., & Bengio , S. (2018) Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems 31*, .

Raghu , M., Gilmer , J., Yosinski , J., & Sohl-Dickstein , J. (2017) Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, .

Maaten , L. & Hinton , G. (2008) Visualizing data using t-sne. *Journal of Machine Learning Research* **9**(86):2579–2605.

Vig , J. & Belinkov , Y. (2019) Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*