Chapter 6: Conclusion

This dissertation has examined how human-centered design principles can be systematically embedded into the modeling of language, particularly through topic modeling and predictive systems. Across its chapters, we have seen the field of natural language processing evolve from static statistical models toward deeply contextual, interactive, and interpretable systems. The overarching goal of this work has been to bridge the divide between interpretability and the expressive power of neural representations, using LLMs where appropriate, while preserving the structure and transparency of probabilistic modeling.

The chapters collectively demonstrate a trajectory from interaction (Chapter 3), to semantic integration (Chapter 5), to human-aligned prediction (Chapter 4). Chapter 2 laid the necessary theoretical foundation by situating this research within the progression of probabilistic, neural, and interactive topic models, and highlighting their respective limitations. Each subsequent chapter then extended these foundations through targeted contributions that progressively achieved the dissertation's central objective: to create topic and language models that are both technically advanced and cognitively interpretable

We address these objectives through a summary of the main contributions (Section 6.1). In the summary, we compare and contrast the findings from each chapter. Finally, we conclude with a discussion of future research directions (Section 6.2).

6.1 Summary of Contributions

The contributions of this dissertation collectively advance the goal of making knowledge discovery through language modeling more adaptive, interpretable, and aligned with human reasoning. Across its studies, the dissertation demonstrates that topic models and large language models need not stand as opposing paradigms, probabilistic models offer interpretability and structure, while neural systems offer contextual precision and generalization. By integrating these strengths, this work lays the foundation for human-in-the-loop knowledge modeling: systems that do not merely generate representations of text, but collaborate with humans in shaping them.

At a technical level, this dissertation made progress on three connected fronts. The first was the creation of interactive neural topic modeling, a framework that transforms topic discovery from a static, one-shot process into an iterative dialogue between human and machine. Traditional topic models assume a fixed generative process, but I-NTM introduced editable topic embeddings that can be continuously updated through minimal user feedback. This shift from symbolic counts to semantic vectors made it possible for users to guide topics with a single constraint, leading to more coherent and human-relevant structures without retraining the entire model. Empirical validation showed that such interaction improves both model coherence and user efficiency in identifying relevant documents, revealing that interpretability can be learned through interaction rather than imposed post hoc.

The second contribution lies in bridging contextual and probabilistic modeling. While the interpretability of topic models is valuable, their lexical rigidity often limits performance on real-world, semantically rich corpora. The Archivist framework introduced in this work addressed this by embedding BERT-derived contextual knowledge directly into Gibbs sampling

as a prior. This integration preserved the transparent inference procedure of probabilistic topic models while grounding them in semantic understanding from pretrained encoders. Experiments demonstrated that Archivist converges faster, produces more coherent topics under smaller data regimes, and yields more stable document–topic representations compared to classical LDA and neural baselines such as CTM and ETM. The result is a new class of models that unify the strengths of both paradigms: interpretable, sample-based inference guided by the expressive capacity of contextual language models.

The third thread extended these modeling ideas into predictive linguistics with large language models. By applying prompting and LoRA fine-tuning to infer psychological dispositions from text, this research examined how LLMs internalize latent constructs such as personality and emotion. These experiments revealed both the potential and the limitations of large models: while Llama achieved lower mean-squared error and more normalized score distributions, their performance exposed the inherent ambiguity of self-reported ground truth. This work thus reframed the role of LLMs, not only as predictors but as instruments for probing the relationship between linguistic expression, model reasoning, and human psychology.

Viewed together, these contributions chart a coherent trajectory toward interactive, context-aware knowledge discovery. From a theoretical perspective, the dissertation demonstrates that probabilistic inference can coexist with neural representation learning when guided by human intent. From an applied standpoint, it provides concrete methodologies, editable embeddings, contextual priors, and fine-tuned LLMs, that can make large-scale text analysis both interpretable and empirically effective.

More broadly, this work contributes to the evolving philosophy of human-centered NLP. It shows that model transparency does not require sacrificing performance, and that interpretability

can emerge from structured interaction rather than architectural simplification. By uniting interaction, contextualization, and inference under one framework, this dissertation advances the state of knowledge discovery: moving from models that describe language to systems that help humans explore, refine, and understand it.

6.2 Future Directions

Section 6.1 summarizes our contributions to our dissertation goal. In this section, we outline a number of directions for future research:

Interactive Neural Topic Models

Currently, users supply arbitrary feedback. A natural extension is to let the model query the user for the most informative feedback. This could be implemented by computing topic uncertainty using entropy over topic–word probabilities or cosine similarity in embedding space, then selecting the most ambiguous topic or word cluster for user input.

Additionally, I-NTM currently re-calculates topic distributions after each label. Future work should allow continuous update streams, incrementally adjusting topic embeddings as users annotate. This could also be in the form of labeling topics throughout the initial model training.

Beyond incremental work, one promising direction based on our results is to embed interactive topic modeling within retrieval-augmented generation (RAG). Here, I-NTM's editable embeddings could act as a controllable interface layer that dynamically restructures retrieval spaces in response to user queries or domain knowledge. This would make topic modeling directly useful for human-in-the-loop information retrieval, systematic review, and agentic systems, areas where static models struggle to adapt to shifting research goals.

Another direction is to integrate interactivity with continual and multimodal learning. Modern corpora evolve rapidly across domains, modalities, and languages. Extending the principles of I-NTM to these settings would involve designing topic representations that evolve online while retaining interpretability, a topic model capable of maintaining coherence even as it absorbs new information.

Looking past algorithmic work, a crucial direction is the human–computer interaction layer. A more comprehensive user interface to log user edits, track how embeddings shift over time, and measure subjective interpretability and efficiency.

Archivist

Sampling under contextual priors is computationally expensive. Future work should explore amortized inference, where a neural network predicts topic assignments given local context and global priors, significantly reducing sampling cost. This could use a variational encoder that initializes sampling with approximate posteriors.

And while topic coherence captures lexical alignment, it does not measure usefulness. Future experiments should expand the evaluative process for Archivist. Measure topic transferability, how stable and informative topics remain when used as features for downstream tasks (RAG, classification, summarization). A systematic evaluation framework for "topic utility" would make Archivist results more interpretable across domains.

Thinking more broadly, future work could extend Archivist's contextual priors to hierarchical or cross-domain topic transfer. A multi-level Archivist could model how high-level topics emerge from domain-specific subtopics, using shared contextual priors across corpora to align knowledge across languages or disciplines. This has immediate applications in digital humanities, and scientific literature analysis, where interpretability and domain transfer are both essential.

Furthermore, Archivist's fusion of contextual embeddings and sampling-based inference suggests a direction for LLM-grounded Bayesian modeling. Instead of using LLMs as black-box generators, they could be incorporated as adaptive prior functions within generative inference frameworks. This would allow models to retain uncertainty estimates, interpretability, and compositional reasoning—properties that purely neural systems often lack. Future work could also explore hybrid architectures where probabilistic inference runs atop LLM-derived embeddings that evolve dynamically as models are fine-tuned or updated, maintaining coherence over time.

LLM Psychological Disposition

A central tenet of future directions is the current evaluative ground-truth. To move beyond noisy self-report labels, future work should integrate domain experts directly into training. A practical path involves collecting textual samples annotated by trained psychologists and fine-tuning via reinforcement learning, where the reward model is based on expert ratings of personality expression rather than numerical self-scores.

In addition, A key limitation of current datasets is their static nature. Future work should build longitudinal corpora where personality expression changes over time (journal entries, social media history). This would allow evaluation of temporal consistency and adaptation, whether models can capture stable traits amid changing language patterns.

Further testing into the predictive power and trasnferability of LLMs should be explored. Follow-up research could explore counterfactual testing: perturbing linguistic markers and observing how predictions shift. This would identify which features drive trait judgments, moving toward explainable and ethically grounded personality inference.

Also, testing whether predictions transfer across modalities (text, audio transcripts) would reveal how personality information is encoded across model families. A reproducible benchmark

suite for cross-model personality prediction would be a valuable community resource.

Each of these future directions reinforces a larger vision: *language modeling as a human-aligned, interpretable process rather than a static prediction task.* Future topic models should dynamically incorporate context, feedback, and expert knowledge; predictive systems should model meaning and intent rather than labels alone. Bridging probabilistic inference, neural embeddings, and LLM reasoning offers a promising path toward systems that understand as well as generate language.

6.3 Closing Statement

Together, these contributions push topic modeling and predictive language research toward systems that are not only technically strong but also interpretable, interactive, and human-aligned. By demonstrating how contextual embeddings, user constraints, and expert-informed fine-tuning can reshape probabilistic and neural models, this work underscores that the next advances in NLP will come not just from larger models, but from centering human needs and agency within the modeling loop.