Thesis Topic:

"Evaluating and Enhancing Figurative Language Understanding in Large Language Models: A Study on Multi-Type Figurative Expressions Using the FLUTE Dataset"

Literature Review

1) Overview of Figurative Language in NLP

Figurative language includes expressions that go beyond their literal meanings, often requiring cultural, contextual, or cognitive understanding to interpret. Common forms include metaphors, idioms, similes, sarcasm, and hyperboles. These constructs play a crucial role in communication, allowing for creativity, persuasion, and emotional expression. However, their inherent complexity makes them a challenging area for computational models to process (Gibbs, 1994). Figurative language plays a significant role in enhancing the depth and richness of both written and spoken communication (Florman, 2017). From a cognitive science perspective, understanding figurative language is critical as it involves abstract thought processes and interpretive adjustments to individual words. This adjustment allows individuals to derive meanings that go beyond the literal sense of words (Michaeli et al., 2008). Michaeli and colleagues highlight the intricate cognitive mechanisms involved in interpreting figurative expressions, emphasizing the relevance of figurative language in both linguistic and psychological studies (Michaeli et al., 2008).

Furthermore, in lexicographical studies, figurative language is viewed as a transference or extension of meaning from its literal sense. This perspective underscores its prevalence in everyday communication, highlighting its role in shaping effective and nuanced language use (Deignan, A., 2015). Springer’s Lexicographical Encyclopedia describes figurative language as an essential component of human interaction, illustrating its integration into both casual and formal language contexts (Hanks, P., de Schryver, GM., 2015). Figurative language enriches human interactions by

allowing speakers to express nuanced ideas and emotions concisely. Various studies emphasize its prevalence in daily communication, literature, and digital media (Lal, Y., & Bastan, M., 2022). For instance, idiomatic expressions can convey cultural meanings that literal language cannot capture, posing challenges for Large Language Models (LLMs) that primarily rely on surface-level text comprehension (Imran, M.M. et al., 2023). This indicates a need for LLMs to be equipped with deeper semantic understanding capabilities to effectively process and generate figurative language.

1.1) Definitions of Figurative Language Forms

Figurative language encompasses various linguistic constructs that go beyond their literal meanings, playing a critical role in enriching communication. Each form of figurative expression contributes uniquely to the depth and nuance of language, enabling the conveyance of abstract ideas, emotions, and cultural nuances. Among these, metaphors stand out as conceptual mappings between domains, allowing one concept to be understood in terms of another. For example, the metaphor “Time is a thief” compares the intangible concept of time to the concrete image of a thief, illustrating the power of metaphor to encapsulate complex relationships (Lakoff & Johnson, 1980). Idioms further illustrate the intricacy of figurative language. These fixed expressions, such as “spill the beans” (reveal a secret) or “kick the bucket” (to die), defy interpretations based on the meanings of their words. Their context-dependent nature makes them particularly challenging for computational models to decode accurately (Katz & Fodor, 1963). Similarly, similes employ explicit comparisons to convey meaning, using connectors like “like” or “as.” Phrases such as “She is as fierce as a lion” directly liken one concept to another, offering a vivid, relatable depiction of bravery (Veale, 2012). Likewise, sarcasm, a more complex form of figurative language, uses irony to mock or convey contempt, often requiring an understanding of tone and intent. Expressions like “Oh great, another rainy day” may mean the opposite of their literal wording, posing interpretive challenges even for advanced language models (Ghosh et al., 2018).

Hyperboles, by contrast, achieve emphasis through intentional exaggeration. Statements such as “I’ve told you a million times” exemplify this form, where the exaggeration underscores the speaker’s frustration or urgency (Colston & O'Brien, 2000).

Other forms, such as personification, metonymy, and synecdoche, add layers of richness to figurative language by imbuing non-human entities with human characteristics, substituting closely associated terms, or allowing parts to represent wholes, respectively. For instance, “The wind whispered through the trees” anthropomorphizes nature to create an evocative image while referring to the White House as the “executive branch” demonstrates metonymy’s utility in communication (Long, 2018; Sun, 2024). Synecdoche operates similarly by focusing on a specific aspect, as seen when the term “sails” is used to represent an entire ship (Brown, 1979). Irony, onomatopoeia, and alliteration further illustrate the versatility of figurative language. Irony conveys meanings opposite to their literal interpretation, often creating unexpected or contradictory outcomes, such as in the scenario of a traffic cop penalized for unpaid parking tickets (Nurdinova et al., 2022). Onomatopoeia, by mimicking sounds, enriches narratives with auditory characteristics, as demonstrated in “The bees buzzed busily among the blooming flowers” (Bredin, 1996). Meanwhile, alliteration enhances the aesthetic appeal of language through the repetition of initial consonant sounds, exemplified in “The swift, silent serpent slithered seamlessly” (Semaeva, 2024).

Puns, or paronomasia, add a playful dimension by exploiting multiple meanings of a term or similar-sounding words for rhetorical or humorous effect. For example, the phrase “I like kids, but I don’t think I could eat a whole one” humorously juxtaposes the meanings of “kid” as both a young child and a young goat, showcasing the interplay of meanings (Giorgadze, 2014).

Together, these forms of figurative language highlight the intricacies of human communication and underline the challenges they present to natural language understanding systems. The ability to process and interpret these constructs is essential for developing effective language models and advancing applications in NLP.

1.2) Importance in NLP

Processing figurative language is crucial in Natural Language Processing (NLP) due to its diverse applications and the complexities it introduces. (Hyewon Jang., et al., 2023) Figurative language includes varieties such as metaphors, similes, idioms, and puns, which convey meanings that extend beyond their literal interpretation. (Bisikalo, O.V., et al., 2019) Understanding these forms of language is fundamental for several key applications in NLP, including sentiment analysis, dialogue systems, and creative text generation.

In sentiment analysis, accurately interpreting figurative language is essential to gauge the true emotional tone behind textual data. (Nguyen, H. L., et al., 2015) For example, a statement like “I'm on cloud nine” suggests extreme happiness rather than a literal interpretation, which could mislead sentiment classification models. (Karamouzas, D., et al., 2022) The presence of figurative expressions can significantly affect the sentiment expressed in a sentence, necessitating models that can recognize and appropriately interpret these nuances. Effective sentiment analysis enhances applications such as customer feedback systems and social media monitoring, allowing for a more accurate understanding of public sentiment. (Tomáš Hercig., et al., 2017)

Dialogue systems, including chatbots and virtual assistants, require a robust understanding of figurative language to engage users effectively. (Karamouzas, D., et al., 2022) In human conversations, figurative expressions are often employed, and failing to recognize these can lead to misunderstandings or inappropriate responses. For instance, if a user states, “It's raining cats and dogs”, the system must interpret this phrase correctly to provide relevant responses instead of a literal interpretation, which could confuse the conversation flow. (Tummala, P., et al., 2024) Developing NLP systems that can process and respond to figurative language helps improve user experience and communication efficacy. (Jhamtani, H., et al., 2021)

Creative text generation, an emerging area within NLP, often incorporates figurative language to produce engaging narratives or generate artistic expressions. Figurative language enriches storytelling by adding depth

and resonance, allowing for more expressive and relatable content. (Repeko, A.P., et al., 2001) For instance, systems designed to generate poetry, or narrative prose must leverage onomatopoeia, metaphors, and other figures of speech to connect with audiences emotionally. (Vulchanova, M.D., et al., 2018) The ability to understand and generate figurative expressions enhances the quality of the texts produced and broadens the potential applications of NLP technology in fields like marketing and entertainment. (Mason, Z.J. 2004)

Despite the importance of processing figurative language, its inherently nonliteral nature poses significant challenges for NLP systems. (Nagels, A., et al., 2013) The disparity between literal and figurative meanings often confounds algorithms, leading to misinterpretations. (Reyes, A. 2013) Many NLP models operate on statistical or machine learning principles that rely heavily on context, but figurative language frequently subverts straightforward contextual cues. This complexity necessitates advanced model architectures, such as those leveraging deep learning techniques, to better capture the subtleties of language use. (Potamias, R.A., et al., 2020)

A significant barrier is the scarcity of annotated datasets tailored to figurative language. Existing datasets often emphasize literal language, creating a training deficit for models handling figurative forms. Without sufficient diversity in training data, models struggle to generalize across different types of figurative expressions. To address these gaps, ongoing research seeks to develop diverse datasets that reflect the complexity of figurative language, enabling models to distinguish effectively between literal and figurative meanings. This work is critical for advancing NLP systems to handle the nuanced and context-dependent nature of figurative communication. (Huiyuan Lai., et al., 2024)

2) Review of Previous Work on Figurative Language Understanding

Research into the understanding of figurative language in NLP has progressed significantly over the years, evolving through four primary paradigms: rule-based systems, statistical models, neural models, and transformer-based models. Each of these approaches has contributed unique

insights and advancements to the field, addressing various challenges posed by the nonliteral and context-dependent nature of figurative expressions.

Early research in figurative language understanding focused on constructing models to decode individual linguistic phenomena. However, as the complexity of figurative expressions became apparent, the need for unified approaches to evaluate linguistic phenomena, including those underlying figurative constructs, gained importance.

Rule-based systems, which rely on manually crafted rules or linguistic patterns, represent some of the earliest efforts in this area. Katz and Fodor (1963) introduced one of the first semantic theories for idioms, highlighting the necessity of noncompositional interpretations. While these systems are interpretable and effective for predefined expressions, they are limited by their inability to scale and adapt to novel or unseen figurative constructs.

The advent of statistical models marked a shift towards probabilistic approaches for figurative language detection. Techniques such as Latent Semantic Analysis (LSA) enabled researchers to model semantic similarities between words and phrases, aiding in the identification of figurative relationships. For instance, Turney et al. (2001) utilized statistical patterns to uncover metaphorical connections by comparing word embeddings within their contextual usage. These models offered improved flexibility over rule-based approaches but still faced challenges in handling complex or deeply nuanced expressions.

The introduction of neural models, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, represented another significant step forward. These models demonstrated promise in detecting figurative language by capturing sequential dependencies within text, a feature critical for understanding idiomatic and metaphorical phrases. As Rei et al. (2017) noted, this ability to process sequential patterns brought greater sophistication to figurative language analysis. However, the limitations of RNNs and LSTMs in capturing long-range dependencies hindered their overall effectiveness, leaving room for further innovation.

Transformer-based architectures, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), have revolutionized the field, leveraging self-attention mechanisms to model global contextual relationships. These advancements have significantly improved the performance of figurative language understanding tasks, including metaphor detection (Liu, J., et al., 2020) and idiom classification (Feldmann et al., 2021). Despite their success, transformer models are not without limitations. They continue to struggle with non-compositionality and the nuanced contextual cues inherent in figurative language, particularly when interpreting sarcasm. This highlights the need for further refinement and innovation to address these challenges fully.

To address these issues, more comprehensive benchmarks have been introduced to evaluate linguistic phenomena systematically. The Holmes benchmark, for instance, offers structured evaluations by isolating linguistic competence from linguistic performance. This methodology enables better assessment of abstract constructs like metaphors and sarcasm, complementing existing approaches in figurative language understanding (Waldis et al., 2024).

The evolution of approaches to figurative language understanding underscores the growing complexity and sophistication of NLP systems, as well as the persistent challenges posed by the abstract and context-dependent nature of figurative expressions. Each paradigm has contributed valuable advancements, paving the way for continued research and development in this critical area.

3) Datasets for Figurative Language Understanding

Datasets are integral to training and evaluating models for figurative language understanding, providing the foundational material that enables computational systems to interpret and process nonliteral expressions. Among the most widely utilized resources is the FLUTE dataset, which is specifically designed for multi-type figurative language understanding. This dataset provides natural language inference (NLI) pairs annotated for entailment and contradiction, encompassing a range of figurative forms such

as idioms, metaphors, sarcasm, and similes (Zheng et al., 2022). Its comprehensive scope makes it a valuable resource for evaluating the performance of models across various figurative types.

The VUA Metaphor Corpus offers another significant contribution by focusing on metaphorical language, particularly verbs. It provides detailed annotations that distinguish between metaphorical and literal meanings, making it an essential tool for understanding the nuanced use of metaphors (Steen et al., 2010). For idiomatic expressions, resources like the Cambridge Idiom Corpus and idiomatic annotations from Wiktionary enable models to capture the contextual subtleties that define idioms. These datasets contribute to the development of systems capable of recognizing and interpreting idiomatic language, which often resists straightforward compositional analysis.

Sarcasm, with its reliance on tone, context, and intent, presents unique challenges for computational models. The Sarcasm Corpus, composed of annotated tweets and comments, facilitates the training of systems to detect sarcasm and understand its often contradictory nature (Ghosh et al., 2018). These datasets collectively address specific facets of figurative language, providing valuable resources for targeted model development.

However, despite their importance, these datasets exhibit notable limitations. Most focus on a single type of figurative expression, such as metaphors, idioms, or sarcasm, which restricts their utility for generalization across multiple figurative forms. This narrow focus often results in models that excel in one domain but fail to adapt effectively to others. The FLUTE dataset stands out in this regard, as its multi-type coverage provides a broader and more balanced foundation for evaluating and enhancing models' comprehensive figurative understanding capabilities. Addressing these limitations through the development of more diverse and inclusive datasets remains a critical avenue for advancing the field of figurative language understanding.

4) Evaluation Methods

Evaluating figurative language understanding in NLP involves a combination of automated and human metrics to comprehensively assess model performance. Automated metrics provide an efficient means of evaluation, with accuracy and F1-score serving as standard measures for classification tasks such as metaphor detection. These metrics assess the precision and recall of models, offering a clear picture of their ability to classify figurative language accurately.

Metrics like BLEU and ROUGE are frequently employed for text similarity tasks, particularly in paraphrasing applications. BLEU (Papineni et al., 2002) evaluates the correspondence between generated text and reference text using n-gram overlap, while ROUGE focuses on recall-oriented measures for the same purpose. Both metrics provide valuable insights into the quality of model-generated text but often fall short in capturing the subtleties of figurative expressions. BERTScore, a more recent contextual embedding-based metric, evaluates semantic similarity by leveraging the embeddings from pretrained language models like BERT (Zhang et al., 2020). This metric offers improved sensitivity to the context and meaning of figurative expressions, making it particularly suitable for evaluating nuanced language use.

Human evaluation, however, remains indispensable for assessing aspects of figurative language that automated metrics often overlook. Human evaluators typically use Likert scales to rate fluency, figurative accuracy, and contextual appropriateness, providing a nuanced understanding of model performance. This approach is especially valuable for evaluating complex constructs like sarcasm and idiomaticity, where contextual understanding and intent play a significant role. For instance, as Ghosh et al. (2018) emphasize, the detection of sarcasm often depends on subtle contextual cues and tone, which are difficult for automated metrics to capture reliably.

By integrating automated metrics with human evaluation, researchers can achieve a more holistic assessment of figurative language understanding. This combination allows for both scalability and depth, ensuring that models

are rigorously evaluated across a spectrum of linguistic challenges. Such comprehensive evaluation frameworks are essential for advancing NLP systems capable of interpreting and generating figurative language with accuracy and contextual sensitivity.

5) Research Gaps

Despite the progress made in figurative language understanding, several significant gaps and challenges persist, limiting the effectiveness and generalizability of current models. One prominent issue lies in the difficulty of generalization across different types of figurative language. Models trained in specific types, such as metaphors or sarcasm, often fail to adapt to other forms of figurative expressions. This specialization limits the broader applicability of these models, underscoring the need for more versatile and comprehensive approaches.

Furthermore, recent studies highlight that large-scale models often struggle with complex phenomena such as reasoning and discourse, which are critical for understanding figurative expressions. These deficiencies are evident even in state-of-the-art benchmarks, revealing gaps in evaluating models' ability to generalize across multiple linguistic phenomena (Waldis et al., 2024). This limitation parallels challenges in figurative language understanding, particularly in capturing contextual subtleties like sarcasm and idiomatic expressions.

Another critical challenge is the requirement for deep contextual reasoning, particularly in understanding sarcasm and idioms. Sarcasm detection often hinges on subtle contextual cues, such as tone and intent, which are not always explicitly available in textual data. Similarly, idiom interpretation demands a nuanced understanding of cultural and linguistic context, further complicating computational processing. Current models frequently struggle to grasp these intricacies, resulting in misinterpretations and reduced performance in real-world applications.

Evaluation methods also present a notable limitation in advancing figurative language understanding. Automated metrics, while efficient and widely used, often fail to capture the creativity and subtlety inherent in figurative expressions. These metrics are typically designed for tasks like classification or text similarity and may not fully reflect the complexities of figurative language. Consequently, there is a pressing need for improved evaluation frameworks that integrate nuanced human judgments with automated methods. Such frameworks would provide a more accurate and holistic assessment of model performance, facilitating the development of systems capable of addressing the multifaceted nature of figurative language.

6) Addressing the Gaps

This thesis aims to address the existing gaps in figurative language understanding by leveraging and extending the capabilities of transformer-based large language models (LLMs) such as ROBERTA, BERT, and T5. These models will be evaluated on the FLUTE dataset, a resource specifically designed for multi-type figurative language understanding. The evaluation will provide insights into the strengths and limitations of current LLMs in handling diverse figurative expressions, including idioms, metaphors, sarcasm, and similes.

To enhance contextual understanding, this research proposes the implementation of advanced techniques such as fine-tuning with additional datasets and innovative preprocessing strategies. These enhancements are intended to equip models with a deeper ability to discern the subtleties of figurative language, particularly in challenging contexts like sarcasm detection and idiom interpretation. In addition to these techniques, recent advancements in evaluation methodologies, such as classifier-based probing as implemented in the Holmes benchmark, provide a structured approach to isolating linguistic competence from linguistic performance (Waldis et al., 2024). Adapting these methods to figurative language processing can enable a more targeted understanding of nuanced constructs like metaphors and sarcasm.

By expanding the range and diversity of training data, the study aims to improve the generalization capabilities of LLMs across different figurative types. Another significant focus of this thesis is the development of robust evaluation frameworks that combine automated metrics with human judgments. While automated metrics offer efficiency and scalability, they often fall short in capturing the creativity and nuanced meanings intrinsic to figurative language.

Additionally, findings from recent work suggest that instruction tuning could play a vital role in aligning LMs with human-like interpretive processes, though further refinement is needed for functional phenomena such as figurative language (Waldis et al., 2024). Incorporating human evaluation into the assessment process ensures a more comprehensive understanding of model performance, particularly in areas where subtle contextual or cultural factors play a critical role.

By addressing these challenges, this study seeks to advance the field of figurative language processing, contributing novel insights into the capabilities and limitations of transformer-based LLMs.

7) Research Question

Based on the identified gaps: "Can fine-tuning on multi-type figurative datasets like FLUTE improve the performance of large language models in understanding and classifying figurative language?"

References

Gibbs, R. W., Jr. (1994). “The Poetics of Mind: Figurative Thought, Language, and Understanding.” Cambridge University Press.

Florman, Ben. (2017). “Figurative Language.” LitCharts. LitCharts LLC, 5 May 2017.

Michaeli et al. (2008). “Figurative Language: “Meaning” is often more than just a sum of the parts.” Association for the Advancement of Artificial Intelligence.

Deignan, A. (2015). “Figurative language and lexicography.” International Handbook of Modern Lexis and Lexicography. Springer, Berlin, Heidelberg.

Hanks, P., de Schryver, GM. (2015). “International Handbook of Modern Lexis and Lexicography.” Springer, Berlin, Heidelberg.

Lal, Y., & Bastan, M. (2022). “SBU Figures It Out: Models Explain Figurative Language.” Proceedings of the 3rd Workshop on Figurative Language Processing (FLP).

Imran, M.M., Chatterjee, P., & Damevski, K. (2023). “Shedding Light on Software Engineering-Specific Metaphors and Idioms.” 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE), 2555-2567.

Lakoff, G., & Johnson, M. (1980). “Metaphors We Live” By. University of Chicago Press.

Katz, J., & Fodor, J. A. (1963). “The Structure of a Semantic Theory.” Language, Vol. 39, No. 2. (Apr. - Jun. 1963), pp. 170-210.

Veale, T. (2012). “Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity.” Bloomsbury Academic.

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018). “Sarcasm analysis using conversation context.” Computational Linguistics, 44(4), 755-792.

Colston, H. L., & O'Brien, J. (2000). “Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole.” Discourse Processes, 30(2), 179–199.

Long, D. (2018). “Meaning Construction of Personification in Discourse Based on Conceptual Integration Theory.” Studies in Literature and Language, 17, 21-28.

Sun, C. (2024). “A Study on the Application of Metonymy in English Poetry.” Scientific and Social Research.

Brown, C.H. (1979). “A Theory of Lexical Change (with Examples from Folk Biology, Human Anatomical Partonomy and Other Domains)” Anthropological Linguistics, 21.

Semaeva, O. (2024). “The Effectiveness of Using Specific English Alliteration in Russian-Language Names and Advertising Slogans.” Linguistics and Intercultural Communication.

Bredin, H. (1996). “Onomatopoeia as a figure and a linguistic principle.” New Literary History, 27(3), 555–569.

Giorgadze, M. (2014). “Linguistic Features of Pun, Its Typology and Classification.” European Scientific Journal, ESJ, 10(10).

Hyewon Jang, Qi Yu, and Diego Frassinelli. (2023). “Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans.” Association for Computational Linguistics: ACL 2023, pages 9816–9832

Bisikalo, O.V., Ivanov, Y.E., & Sholota, V. (2019). “Modeling the Phenomenological Concepts for Figurative Processing of Natural-Language Constructions.” International Conference on Computational Linguistics and Intelligent Systems.

Nguyen, H. L., Trung, D. N., Hwang, D., & Jung, J. J. (2015). “KELabTeam: A Statistical Approach on Figurative Language Sentiment Analysis in Twitter.” International Workshop on Semantic Evaluation.

Karamouzas, D., Mademlis, I., & Pitas, I. (2022). “Neural Knowledge Transfer for Sentiment Analysis in Texts with Figurative Language.” 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP).

Tomáš Hercig and Ladislav Lenc. (2017). “The Impact of Figurative Language on Sentiment Analysis.” The International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 301–308, Varna, Bulgaria. INCOMA Ltd.

Karamouzas, D., Mademlis, I., & Pitas, I. (2022). “Neural Knowledge Transfer for Sentiment Analysis in Texts with Figurative Language.” 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP).

Tummala, P., & Roa, C.K. (2024). “Exploring T5 and RGAN for Enhanced Sarcasm Generation in NLP.” IEEE Access, 12, 88642-88657.

Jhamtani, H., Gangal, V., Hovy, E., & Berg-Kirkpatrick, T. (2021). “Investigating Robustness of Dialog Models to Popular Figurative Language Constructs.” Conference on Empirical Methods in Natural Language Processing.

Repeko, A.P., Radchikova, N.P. (2001). “Natural Language Understanding: Problems of Figurative Language Processing.” Belarusan Foundation of Fundamental Research.

Vulchanova, M.D., & Vulchanov, V. (2018). “Figurative language processing: A developmental and NLP Perspective.” CLIB.

Mason, Z.J. (2004). “CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System.” Association for Computational Linguistics, 30, 23-44.

Nagels, A., Kauschke, C., Schrauf, J., Whitney, C., Straube, B., & Kircher, T. (2013). “Neural substrates of figurative language during natural speech perception: an fMRI study.” Frontiers in Behavioral Neuroscience.

Reyes, A. (2013). “Linguistic-based Patterns for Figurative Language Processing: The Case of Humor Recognition and Irony Detection.” Proces. del Leng. Natural, 50, 107-109.

Potamias, R.A., Siolas, G. & Stafylopatis, A. (2020). “A transformer-based approach to irony and sarcasm detection.” Neural Comput & Applic 32, 17309–17320

Huiyuan Lai and Malvina Nissim. (2024). “A Survey on Automatic Generation of Figurative Language: From Rule-based Systems to Large Language Models.” ACM Comput. Surv. 56, 10, Article 244.

Nurdinova, G. Sh., & Egamnazarova, D. Sh. (2022). “Irony as a Multipurpose Stylistic Device.” Current Research Journal of Philological Sciences.

Brown, T., et al. (2020). “Language Models are Few-Shot Learners.” NeurIPS.

Devlin, J., et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT.

Feldman, N., et al. (2021). “Analyzing Idiomatic and Literal Usage with Pre-trained Language Models.” ACL.

Rei, M., et al. (2017). “Jointly Learning Semantic Parser and Natural Language Generator.” EMNLP.

Papineni, K., et al. (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation.”

Liu, J., et al. (2020). “Metaphor Detection Using Contextual Word Embeddings from Transformers.” ACL.

Turney, P. D., et al. (2001). “Mining the Web for Synonyms.” Information Retrieval Journal.

Andreas Waldis, et al. (2024). “Holmes ⌕ A Benchmark to Assess the Linguistic Competence of Language Models.” Association for Computational Linguistics 2024; 12 1616–1647.

Zhang, T., et al. (2020). “BERTScore: Evaluating Text Generation with BERT.” ICLR.