Large Language and (Vision) Models often focus on literal interpretations of text and images. However, much of human communication involves the creative use of language or visuals. Automatic language and vision generation tools have the potential to assist human writers by suggesting creative phrasings, plot twists, or even entirely novel narrative arcs, catalyzing a synergy between human and machine creativity. From adaptive e-books that modify narratives based on a reader's preferences, to personalized advertisement campaigns using visual metaphors that resonate with a particular audience, creative text/vision generation can offer bespoke content at scale. Additionally, creative text such as humor, similes, metaphors, or idioms often carry cultural connotations; by building computational models that understand or generate such creative text, we can facilitate more effective culture-aware dialog systems.

One of the salient challenges for generating creative content stems from the inherent nature of creativity, which demands a blend of expansive common sense knowledge and the capacity for divergent thinking (Baer, 2014). Traditional text or image understanding and generation paradigms are heavily reliant on vast volumes of training data. Acquiring meticulously annotated data for creative endeavors is not only resource-intensive but also challenging due to the scarcity of qualified annotators. In contexts where such task-specific data is crucial, simply having a model learn from and replicate broader distributions from the raw data used in large language models (LLMs) pre-training can fall short. Genuine creativity often entails deviation from the established norm. However, the primary goal of LLM pre-training is to capture and reproduce this norm, which may not always align with the nuances of authentic creative innovation. Furthermore, it is worth noting that the evaluation of many natural language processing (NLP) models is conducted in siloed environments with input from non-expert crowd workers. This approach can potentially hinder their efficacy in interactive scenarios with domain experts. In the realm of creative tasks, it becomes imperative for these models not only to comprehend the intricacies of expert requirements but also to adapt, assist, and enhance their expertise progressively. My research seeks to build machine-learning models for creativity by equipping them with **implicit/commonsense knowledge** as well as developing **human-centered robust evaluation frameworks in both standalone or interactive settings by relying on technical skills from computer science and design in combination with other disciplines, including the humanities**. Toward this goal, I have made targeted contributions across three broad topics:

1. *Knowledge Enhanced Models for Creative Text Generation* I have demonstrated the utility of grounded commonsense knowledge in generating and understanding figurative language such as sarcasm, simile, metaphors and idioms (Chakrabarty et al., 2020b,a, 2021; Stowe et al., 2021; Chakrabarty et al., 2022a).

2. *Human AI collaboration frameworks for creative tasks across different modalities* My research has led to the design and development of datasets to improve the explainability of creative language understanding models (Chakrabarty et al., 2022b) using human-AI collaborative framework. In addition, I have built a creative writing support tool that is grounded in the theoretical cognitive process model of writing (Chakrabarty et al., 2023b). Aside from text, I have also built methods and tools to allow illustrators to depict visual metaphors through collaboration between large language models and vision and language models (Chakrabarty et al., 2023c).

3. *Evaluation framework for creative writing* My research has integrated theoretical foundations from cognitive sciences (Torrance, 1966) and social psychology (Amabile, 1982) leading to the development of a rubric for Creative Writing evaluation (Chakrabarty et al., 2023a) that has been used by domain experts to discriminate and identify AI written text from that of expert-written text. Our evaluation framework highlights the limitations of large language models in producing creative text as well as their inability to assess creative text, thereby enabling the room to build better models that align with human-level creativity.

1 Knowledge Enhanced Models for Creative Text Generation

For creativity in particular, commonsense knowledge acts as a foundational scaffold from which truly inventive and impactful deviations can emerge. Without a grasp of commonsense, the combinations might be entirely arbitrary, reducing the chance of producing meaningful or resonant creative output. Additionally, playing with or subverting commonsense expectations can be a source of humor, surprise, or insight, which are central to many creative endeavors. To counter the lack of naturally occurring training data, my research has tackled both unsupervised and weakly supervised knowledge-enhanced methods for generating figurative language such as sarcasm, simile, and metaphors. At the crux of these methods, we augment LLMs with commonsense knowledge generated by knowledge models such as COMET (Bosselut et al., 2019).

Sarcasm Generation: My ACL 2020 paper on Sarcasm Generation Chakrabarty et al. (2020a) takes a non-sarcastic input and converts it into a sarcastic output using an *unsupervised* retrieval-based approach that is primarily guided by two *theoretically-grounded factors*: reversal of valance and incongruity with context that can contain commonsense knowledge. Thus, our approach first replaces negative evaluative words in non-sarcastic messages with their antonyms and then uses concept-centric commonsense knowledge from COMET to generate additional context. Once a commonsense concept is identified, relevant sentences are retrieved from an online dictionary containing that concept. Lastly, to ensure the context sentences are incongruous with the initial non-sarcastic message, they are ranked using a RoBERTa-large model trained on an NLI task, with contradiction scores guiding the selection (See Figure 1).

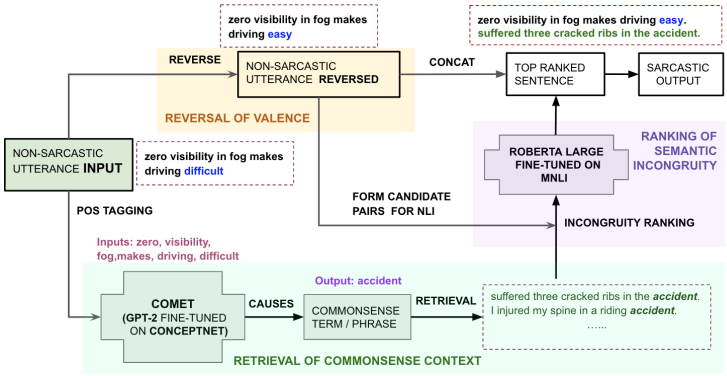

Figure 1: Complete pipeline for sarcasm generation.

Simile and Metaphor Generation: To generate figurative language from its literal counterpart we need large-scale parallel data to fine-tune any existing generative model. However such data is expensive to collect and annotate. To tackle this in my EMNLP 2020 and NAACL 2021 papers on generating similes and metaphors from literal sentences, I utilize external knowledge from commonsense models such as COMET along with insights from linguistic theories to create parallel data. This data is then used for fine-tuning pre-trained seq2seq models such as BART (Lewis et al., 2020) for [literal → figurative] text generation.

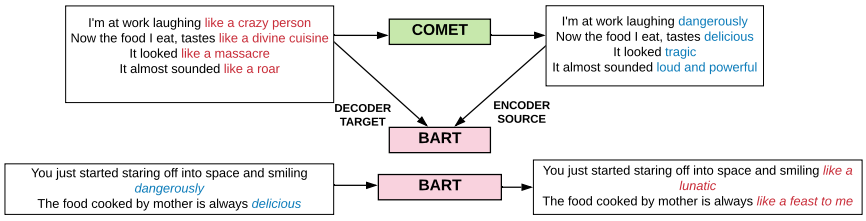

Figure 2: A schematic illustration of our simile generation system, where the top block shows our **training** process while the block below shows the **inference** step.

A simile is usually made of a `TOPIC` and a `VEHICLE`, both of which are typically noun phrases compared implicitly/explicitly via a shared `PROPERTY`. In my EMNLP 2020 paper (Chakrabarty et al., 2020b) on generating similes I used *HasProperty* relation from COMET to create parallel data (See Figure 2) and finetuned a seq2seq model on that data. Experiments show that our approach generates 88% novel similes that do not share properties with the training data and human judges consistently prefer generations from our system to be more creative and relevant over several compelling baselines.

Similarly in my NAACL 2021 paper (Chakrabarty et al., 2021) I tackled the problem of generating verbal metaphors by first creating a corpus of literal/metaphoric paraphrases. Towards this, we first converted sentences containing metaphorical verbs to their literal counterparts using a Masked Language Model and further used the *SymbolOf* relation from COMET to ensure semantic consistency. This parallel data is used further to finetune a seq2seq model. Human evaluation on an independent test set of literal statements shows that our best model generates better metaphors than three well-crafted baselines 66% of the time on average. In follow-up work at ACL 2021 (Stowe et al., 2021) we reuse this data and extend it to generate conceptual metaphors based on Conceptual Metaphor Theory (Lakoff and Johnson, 1980).

Interpreting Idioms and Simile in Longer Context: In my TACL 2022 paper (Chakrabarty et al., 2022a) I tackle the novel task of how well can language models interpret figurative languages such as idioms and similes. Towards this given a 5 sentence narrative ending in an idiom or simile, I proposed the task of generating a plausible next sentence that is coherent with the context and consistent with the meaning of the figurative expression. Motivated by how L2 learners comprehend unknown figurative language (Cooper, 1999), for narratives ending with an idiom we augmented the last sentence of the narrative with discourse-aware commonsense knowledge (Gabriel et al., 2021) from its context, while for narratives ending with a simile we augmented the last sentence of the narrative with concept-centric commonsense from COMET. Our experiments show that models based solely on pre-trained language models perform substantially worse than humans on these tasks and that knowledge-enhanced models, adopting human strategies for interpreting figurative language types — inferring meaning from the context,

and/or relying on the constituent words' literal meanings improve the performance significantly further bridging the gap from human performance.

2 Human AI collaboration frameworks for creative tasks across different modalities

Model-in-the-loop approaches (i.e., GPT-3 Brown et al. (2020) and crowdsourcing) have been recently proposed to generate datasets, as well as free-form textual explanations (a.k.a natural language explanations Camburu et al. (2018)) for model decisions Liu et al. (2022); Wiegreffe et al. (2021). In my EMNLP 2022 paper FLUTE Chakrabarty et al. (2022b), we tackle the task of understanding figurative languages such as sarcasm, idioms, similes, and metaphor via the lens of textual entailment with natural language explanations justifying label prediction. Our data consists of 9000 <premise(literal), hypothesis(figurative)> pairs created via effective collaboration between GPT3 and crowdworkers. My work shows the power of human-AI collaboration in creating high-quality datasets for complex tasks for which there might be a shortage of skilled annotators. Additionally, this allows distillation into smaller models that can both predict and explain. We also introduce a new evaluation metric that not only accounts for model prediction accuracy but also explanation correctness.

In my ACL 2023 paper, Chakrabarty et al. (2023c) I propose a new task of generating visual metaphors from linguistic metaphors. This is a challenging task for diffusion-based text-to-image models, such as DALL·E 2 Ramesh et al. (2022) since it requires the ability to model implicit meaning and compositionality. I propose to solve the task through the collaboration between Large Language Models (LLMs) and Diffusion Models: Instruct GPT-3 (davinci-002) with Chain-of-Thought prompting generates text that represents a visual elaboration of the linguistic metaphor containing the implicit meaning and relevant objects, which is then used as input to the diffusion-based text-to-image models. Using a human-AI collaboration framework, where humans interact both with the LLM and the top-performing diffusion model, we create a high-quality dataset containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. Evaluation by professional illustrators shows the promise of LLM-Diffusion Model collaboration for this task.

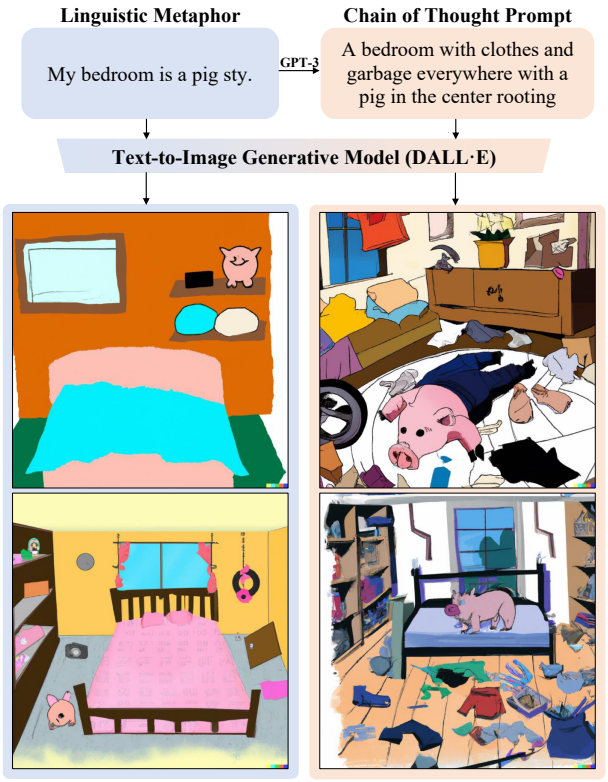

Figure 3: Visual metaphors generated by DALL·E 2 for the linguistic metaphor "My bedroom is a pig sty".

In recent work (Chakrabarty et al., 2023b) under submission I designed a human-AI collaboration framework for writing support [1] that is *designed based on the cognitive process model of writing* (Flower and Hayes, 1981), unlike prior work (Mirowski et al., 2023; Ippolito et al., 2022), and that allows a professional writer to seek help from a LLM (GPT-3.5) *during all three cognitive activities — planning, translating and reviewing —, in a non-linear fashion.* We observe that while the writers use the LLM for all stages of creative writing, they find the model most helpful for *translation-based* subtasks such as targeted rewriting of paragraphs in the text, or *review-based* subtasks such as obtaining feedback on their draft. We also qualitatively analyze the post-completion survey feedback provided by the writers to identify the model's strengths and weaknesses. Current models are limiting for professional writers in several ways including their repetitive nature, over-reliance on clichés and tropes, lack of nuance, subtext, or symbolism as well as overly moralistic and predictable endings. While some of these are perhaps actionable weaknesses that could be resolved with better prompting, our writers also highlight broader concerns pertaining to the model's inability to generate text related to darker topics, as well as difficulty in understanding the writer's intent. On the other hand, writers find the model to be the most helpful as a rewriting tool or feedback provider rather than an original idea generator.

[1] https://collab-stories.github.io/

3 Evaluation framework for creative writing

In recent work (Chakrabarty et al., 2023a) we adapt the Torrance Test for Creative Thinking (TTCT), a protocol for evaluating creativity as a process, and align it for the evaluation of creativity as a product particularly focusing on short stories. Using the Consensual Assessment Technique (Amabile, 1982), which states that the most valid assessment of the creativity of an idea or creation in any field is the collective judgment of experts in that field, we design 14 tests in collaboration with experts called the *Torrance Test for Creative Writing (TTCW)* based on the four original Torrance dimensions of *Fluency, Flexibility, Originality, and Elaboration*. We experimentally validate the TTCW through an assessment of 48 stories involving 10 participants with expertise in creative writing, finding that they reach moderate agreement when administering individual tests, and strong agreement when evaluating all tests in aggregate. We study the abilities of LLMs to generate stories that pass/fail the TTCW tests and their ability to reliably assess the creativity of stories following the TTCW framework through correlation with human judgments. Our findings show that LLM-generated stories are three to ten times less likely to pass TTCW tests compared to expert-written stories, as well as the fact that current state-of-the-art LLMs are not yet capable of reproducing expert assessments when administering TTCW tests. To enable future research in this fast-evolving domain, we release the large-scale annotation of 2,000+ TTCW assessments[2], each accompanied by a natural language expert explanation. Finally, we also discuss how creative writing experts can distinguish between AI vs. human written stories and how future work can use our evaluation framework for building rich interactive writing support tools.

4 Future Directions

As an assistant professor, I will continue my work on understanding the limitations and potentials of neural language models and methods for developing trustworthy and impactful socio-technical systems, building collaborations with researchers in Human-Computer Interaction, Computer Vision, and Machine Learning. I also plan to maintain and grow my collaborations with industry partners as well as other departments such as Creative Writing, Cognitive Sciences, and Psychology.

Improving model performance through better quality data: I aim to investigate how AI-in-the-loop interactive systems can accelerate and improve human annotation as in my prior work (Chakrabarty et al., 2022b, 2023c). In NLP, a lot of dataset collection is done by crowdsourced non-expert individuals who get paid for each annotation. This approach may not encourage careful work. Karpinska et al. (2021) have shown how recruiting high-quality experts leads to better dataset collection and model evaluation for open-ended NLG tasks. I am excited to continue along this line of work. As the models get bigger and better the amount of data required for supervision will likely decrease requiring emphasis on quality over quantity.

Build better evaluation frameworks guided by theoretical work The LLM research community has proposed meta-benchmarks such as BigBench (Srivastava et al., 2022), GMMSK (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020) to standardize evaluation and benchmarking. The multifaceted potential of LLMs and their performance on these benchmarks is thrilling, however, they tend to push towards standardization by fueling various applications with one commonly termed "all-purpose" model. Liao and Xiao (2023) re-frame the goal of developing model evaluation methods as narrowing the socio-technical gap along two dimensions: context realism and human requirement realism, each having possible trade-offs with pragmatic costs to conduct the evaluation. Toward this, I want to follow my prior work on conducting evaluations (Chakrabarty et al., 2023b,a), especially for NLG with domain experts in both standalone and collaborative settings that are designed based on theoretical grounding from linguistics, psychology, and cognitive sciences.

Understanding the impact of RLHF on LLM performance Language models (LMs) often exhibit undesirable text generation behaviors, including generating false, toxic, or irrelevant outputs. Reinforcement learning from human feedback (RLHF) — where human preference judgments on LM outputs are transformed into a learning signal — has recently shown promise in addressing these issues. While these techniques help make models more aligned with human values, they can be a bottleneck for open-ended writing tasks. Padmakumar and He (2023) show that models that are trained with reinforcement learning from human feedback lead to a statistically significant reduction in diversity. My recent work (Chakrabarty et al., 2023b) had similar findings where I show that these models are not perfectly aligned with the goals of a creative writing assistant and often lead to overtly moralistic outputs. Towards this, I plan to construct pre-trained models better aligned to the desired values of creative writers that enable controlled risk-taking, and letting users define the limits of suitable behavior during prediction. Additionally, I also plan to investigate if LMs trained with reinforcement learning on expert preferences lead to improved performance compared to those trained with non-expert human preferences.

[2] https://github.com/salesforce/creativity_eval

References

Teresa M Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997.

John Baer. 2014. *Creativity and divergent thinking: A task-specific approach*. Psychology Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020a. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023a. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020b. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023b. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023c. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Thomas C. Cooper. 1999. Processing of idioms by L2 learners of english. *TESOL Quarterly*, 33(2):233–262.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz12, Ronan Le Bras, Maxwell Forbes12, and Yejin Choi12. 2021. Paragraph-level commonsense transformers with recurrent memory.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago and London.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Ellis Paul Torrance. 1966. *Torrance tests of creative thinking: Norms-technical manual: Verbal tests, forms a and b: Figural tests, forms a and b*. Personal Press, Incorporated.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.