Statement of Purpose

I aim to have a career where I can pursue my research interests, teach, and mentor aspiring researchers. To that end, I plan to become a professor. My research interests include **applied mathematics** (primarily topology and transportation theory), **machine learning** (in particular, unsupervised and efficient methods), **natural language processing** (NLP), and **historical linguistics**. As a master's student, advised by Prof. Taylor Berg-Kirkpatrick, my work focuses on applying computational methods to various domains, such as creative writing, paleography, and numeracy.

Conversation infilling with large language models. My most recent publication [27] (accepted to EMNLP 2022) lies at the intersection of controllable language generation and creative writing. Specifically, I propose and explore the cooperative generation of conversations, dubbed *conversation infilling*. This task aims to generate a seamless bridge of utterances connecting a given pair of source and target utterances. Conversation infilling finds relevance in the film and video game industries, where writers typically compose numerous dialogues between characters. Assisting these writers with large language models may enable higher throughput and spur ideas and creativity, similar to DALL-E [29] for illustrators and Copilot [20] for programmers.

Inspired by NeuroLogic A*esque [22], my proposed approach for this task is Heuristic Guided Lookahead Decoding (HeLo). This decoding strategy does not require finetuning or additional models outside the generating model itself. Instead, before committing to any token, HeLo performs greedy lookaheads to generate *potential* future conversations. Then, among these possible futures, HeLo prioritizes the future state (and its associated token) that brings the conversation closer to the target utterance with a heuristic scoring function. When paired with a transformer trained on open-ended dialogue, quantitative and qualitative evaluations suggest HeLo is a promising method. Regarding my contributions to this publication, I proposed both the task (conversation infilling) and approach (HeLo), performed all development work, organized the human evaluations, and wrote the paper.

Beyond developing approaches for AI-assisted writing, I am particularly interested in exploring other efficient, controllable language generation methods that do not require finetuning. Given the rapid growth of language model size, even the pre-train and finetune paradigm may soon be insufficient for academic labs and independent researchers to study the next iteration of large language models. Prompt-based control methods [21] and decoding strategies similar to HeLo may offer a partial solution, making them worthwhile to study.

Enciphered language identification and decipherment with the Weisfeiler-Lehman distance. My current work lies at the intersection of graph theory, optimal transport, and paleography; it is a collaboration with my advisor and Prof. Yusu Wang. I explore the question: can modern computational methods decipher ancient undeciphered writing systems? Examples of such systems include the Cypro-Minoan syllabary of Cyprus and the Ba-Shu scripts of southwestern China. The first step in decipherment is language identification, a well-studied task for *known* writing systems (e.g., Latin and Devanagari) commonly solved using supervision. However, no such supervision exists in the case of undeciphered writing systems, necessitating the exploration of unsupervised techniques.

Consequently, I turn to the fields of optimal transport and graph theory, where the Weisfeiler-Lehman (WL) distance [24] was recently introduced to quantify the degree to which two labeled graphs (specifically, labeled measure Markov chains) fail the WL isomorphism test [1]. I propose that measuring the WL distance between character-level n-gram language models (encoded as graphs) is analogous to measuring the linguistic distance between the models' training text. Then, given a sample of an undeciphered writing system and a set of candidate languages, the candidate nearest to the sample is most likely its language.

To investigate the soundness of my proposal, I explore the simplified task of identifying the language of known writing systems encrypted with a monoalphabetic simple substitution cipher. Quantitative experiments show that my proposed method outperforms many in the literature [12] and is robust to noise and out-of-domain settings. Moreover, because the Wasserstein distance computation is a subprocess of the WL distance computation, the optimal transport matrix between the language

model nodes function as a cryptographic key which I use to crack the enciphered text with near-perfect accuracy. I am preparing these results for publication (ACL 2023) and currently investigating ways to apply the technique to more difficult ciphers (homophonic and polyalphabetic) and ancient undeciphered writing systems in collaboration with historical linguist Dr. Christina Skelton.

Given its theoretical justifications and simplicity, I find methods such as the WL distance particularly appealing. Moreover, despite the effectiveness and popularity of neural methods, it is refreshing to work with its alternatives. In future work, I plan to continue investigating ways to apply lesser-known, theoretically justified algorithms to NLP tasks.

Numeracy for large language models. My first publication [30] (accepted to NAACL 2022) studies numeracy for language models: the ability of language models to understand and use numbers. For example, consider the statements "The dog weighs 400 pounds." and "The dog weighs 40 pounds." Clearly, the prior is less plausible than the latter, and a numerate language model's learned distribution should reflect this. Unfortunately, most language models treat numbers as another word within their fixed vocabulary. By ignoring the continuous nature of numbers, most are out-of-vocabulary and, therefore, inaccessible to the language model.

While prior work in numeracy exists, my co-authors and I investigate *measurements*: a particular class of numbers with a well-defined and typed system of units. For example, in the statement, "Alex Honnold climbed for 1,000 meters.", the number is "1,000", the unit is "meters," and the dimension is length. Together, they compose a measurement. A better understanding of measurements may benefit downstream applications such as question answering and information extraction, particularly concerning scientific documents. To evaluate language models' ability to reason about measurements, we propose the masked measurement prediction (MMP) task: given some text with its measurement removed, predict both the number and its unit. We demonstrate that most pretrained language models fail at this task. Finetuning these models on MMP offers modest gains, but we show that a custom model that jointly learns to predict the dimension, unit, and number (by parameterizing the location of a Log-Laplace distribution) achieves the best results. Furthermore, our proposed model outperforms several human annotators on MMP, reinforcing the task's difficulty. Regarding my contributions to this publication, I performed much of the development work, experiment management, and writing.

Potential advisors at Stanford. While I am very much open to working with any faculty who study NLP, I am most familiar with the work of Profs. **Dan Jurafsky** and **Tatsunori Hashimoto**. Dan Jurafsky and I share an interest in NLP and its application to the social sciences. In particular, his work on semantic change [10, 11, 14], the language of food [9, 13], and Cantonese [2] is of interest to me. Tatsunori Hashimoto's interest in natural language generation [17] also aligns with my own. Specifically, I find his work on its evaluation [15, 25] and control [28] appealing.

Potential advisors at UC Berkeley. While I am very much open to working with any faculty who study NLP, I am most familiar with the work of Profs. **Dan Klein** and **David Bamman**. Dan Klein and I share an interest in unsupervised learning and historical linguistics. In particular, his work on decipherment [5, 7] and the reconstruction of ancient languages [3, 8, 26] is of interest to me. David Bamman's interest in NLP and its application to the social sciences also aligns with my own. Specifically, I find his work on fiction [16, 18] and historical documents [4, 6] appealing.

Potential advisors at USC. While I am very much open to working with any faculty who study NLP, I am most familiar with the work of Profs. **Jonathan May** and **todo**. Jonathan May and I share an interest in [] and []. In particular, his work on decipherment [19, 23] and todo [**empty citation**] is of interest to me. [todo]'s interest in NLP and its application to the social sciences also aligns with my own. Specifically, I find his work on fiction [16, 18] and historical documents [4, 6] appealing.

Potential advisors at UCLA.

Potential advisors at UCSB.

Potential advisors at UCI.

Potential advisors at UW.

Potential advisors at UCSD.
Potential advisors at UCSC.

References

[1] Boris Weisfeiler and Lehman. "A Reduction of a Graph to a Canonical Form and an Algebra Arising during This Reduction". In: *Nauchno-Technicheskaya Informatsiya*. 1968. URL: https://www.iti.zcu.cz/wl2018/pdf/wl_paper_translation.pdf.

[2] Dan Jurafsky. "On the semantics of cantonese changed tone or women, matches, and chinese broccoli". In: *Annual Meeting of the Berkeley Linguistics Society*. Vol. 14. 1988, pp. 304–318.

[3] Alexandre Bouchard-Côté et al. "A probabilistic approach to diachronic phonology". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 2007, pp. 887–896.

[4] David Bamman and Gregory Crane. "Measuring historical word sense variation". In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 2011, pp. 1–10.

[5] Taylor Berg-Kirkpatrick and Dan Klein. "Simple Effective Decipherment via Combinatorial Optimization". In: *EMNLP*. 2011.

[6] David Bamman, Adam Anderson, and Noah A Smith. "Inferring social rank in an old assyrian trade network". In: *arXiv preprint arXiv:1303.2873* (2013).

[7] Taylor Berg-Kirkpatrick and Dan Klein. "Decipherment with a Million Random Restarts". In: *EMNLP*. 2013.

[8] Alexandre Bouchard-Côté et al. "Automated reconstruction of ancient languages using probabilistic models of sound change". In: *Proceedings of the National Academy of Sciences* 110 (2013), pp. 4224–4229.

[9] Dan Jurafsky. *The language of food: A linguist reads the menu*. WW Norton & Company, 2014.

[10] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2116–2121. DOI: 10.18653/v1/D16-1229. URL: https://aclanthology.org/D16-1229.

[11] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1489–1501. DOI: 10.18653/v1/P16-1141. URL: https://aclanthology.org/P16-1141.

[12] Bradley Hauer and Grzegorz Kondrak. "Decoding Anagrammed Texts Written in an Unknown Language and Script". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 75–86. DOI: 10.1162/tacl_a_00084. URL: https://aclanthology.org/Q16-1006.

[13] Dan Jurafsky et al. "Linguistic markers of status in food culture: Bourdieu's distinction in a menu corpus". In: *Journal of Cultural Analytics* 1.1 (2016), p. 11064.

[14] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (Apr. 2018). DOI: 10.1073/pnas.1720347115. URL: https://doi.org/10.1073%2Fpnas.1720347115.

[15] Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. *Unifying Human and Statistical Evaluation for Natural Language Generation*. 2019. DOI: 10.48550/ARXIV.1904.02792. URL: https://arxiv.org/abs/1904.02792.

[16] David Bamman. "Litbank: Born-literary natural language processing". In: *Computational Humanites, Debates in Digital Humanities (2020, preprint)* (2020).

[17] Daniel Kang and Tatsunori Hashimoto. "Improved natural language generation via loss truncation". In: *arXiv preprint arXiv:2004.14589* (2020).

[18] Matthew Sims and David Bamman. "Measuring Information Propagation in Literary Social Networks". In: *EMNLP*. 2020.

[19] Nada Aldarrab and Jonathan May. "Can Sequence-to-Sequence Models Crack Substitution Ciphers?" 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 7226–7235. DOI: 10.18653/v1/2021.acl-long.561. URL: https://aclanthology.org/2021.acl-long.561.

[20] Mark Chen et al. *Evaluating Large Language Models Trained on Code*. 2021. DOI: 10.48550/ARXIV.2107.03374. URL: https://arxiv.org/abs/2107.03374.

[21] Pengfei Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. 2021. DOI: 10.48550/ARXIV.2107.13586. URL: https://arxiv.org/abs/2107.13586.

[22] Ximing Lu et al. *NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics*. 2021. DOI: 10.48550/ARXIV.2112.08726. URL: https://arxiv.org/abs/2112.08726.

[23] Nada Aldarrab and Jonathan May. "Segmenting Numerical Substitution Ciphers". In: *arXiv preprint arXiv:2205.12527* (2022).

[24] Samantha Chen et al. "Weisfeiler-Lehman meets Gromov-Wasserstein". In: arXiv, 2022. DOI: 10.48550/ARXIV.2202.02495. URL: https://arxiv.org/abs/2202.02495.

[25] Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. "Spurious Correlations in Reference-Free Evaluation of Text Generation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1443–1454. DOI: 10.18653/v1/2022.acl-long.102. URL: https://aclanthology.org/2022.acl-long.102.

[26] Andre He, Nicholas Tomlin, and Dan Klein. *Neural Unsupervised Reconstruction of Protolanguage Word Forms*. 2022. DOI: 10.48550/ARXIV.2211.08684. URL: https://arxiv.org/abs/2211.08684.

[27] Ivan Lee and Taylor Berg-Kirkpatrick. "HeLo: Learning-Free Lookahead Decoding for Conversation Infilling". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022. URL: https://drive.google.com/file/d/15c5fEpjAbeaRnRcVrdoYIjq9VM_KLCyw/view?usp=sharing.

[28] Xiang Lisa Li et al. *Diffusion-LM Improves Controllable Text Generation*. 2022. DOI: 10.48550/ARXIV.2205.14217. URL: https://arxiv.org/abs/2205.14217.

[29] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. DOI: 10.48550/ARXIV.2204.06125. URL: https://arxiv.org/abs/2204.06125.

[30] Daniel Spokoyny et al. "Masked Measurement Prediction: Learning to Jointly Predict Quantities and Units from Textual Context". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 17–29. DOI: 10.18653/v1/2022.findings-naacl.2. URL: https://aclanthology.org/2022.findings-naacl.2.