Edge of Knowledge: Probing Language Model Boundaries to Mitigate Hallucinations

Motivation: A recent study examined an alarming behavior of code-generating Large Language Models (LLMs): package hallucinations[1]. These occur when an LLM generates fake package names in code suggestions. One danger of this behavior is package confusion attacks, whereby adversaries publish malicious code with the same name as the hallucinated package. The study found over 200,000 unique examples, highlighting how prevalent this behavior has become. This is just one example of how LLM hallucinations can generate inaccurate or misleading information at scale, posing a significant risk to society. Hallucinations, when a model generates plausible-looking content that is actually nonsensical or factually incorrect, can be categorized in a variety of different ways[2]. In this work, I focus specifically on instances of factual inconsistency: when the model's outputs contradict real world facts, suggesting that the model is operating at its knowledge boundary. In other words, when the model begins "lying" because it doesn't "know" something.

Currently, there are many different strategies to detect and curb hallucinatory behavior, including retrieval-augmented generation[3], black-box prompting methods[4], and computing the semantic entropy of output logits[5]. These approaches do not take into account internal model behavior, which may have adverse effects due to the demonstrated discrepancy between model output and internal activations[6]. Alternatively, a white-box approach involves training classifier probes to predict semantic entropy from internal model states[7]. Kossen et al. found that when semantic entropy is high, it is more likely that a hallucination may occur because the language model is experiencing greater uncertainty. This research did not, however, attempt to mitigate the behavior and the investigation of internal model structures was minimal. Overall, generalizable hallucination detection and mitigation remains an open challenge in language modeling because of the inherent complexity of models and their tasks.

Hypothesis: I hypothesize that hallucinations in language models can be more thoroughly detected and mitigated by ***1) pinpointing model knowledge boundaries through the identification of internal substructures related to hallucinations***, and ***2) developing a novel mitigation method that alters the weights of these substructures, modifying model behavior when knowledge boundaries are hit.*** The core research question is: *can a language model tell you when it's unsure?*

Aim 1: Detect neuronal substructures related to hallucinations via semantic entropy probing. A key focus at this step is identifying when the model hits a knowledge boundary. This boundary is where factual inconsistencies occur: the model is highly uncertain about the correct prediction, resulting in mistakes. By taking a subset of the Natural Questions corpus[8], I will construct a handpicked test set that spans the model's knowledge range in a specific domain such as Wikipedia-style information about a historical figure. This will give insight into when the model begins hallucinating as a function of the data it is prompted by.

I will begin by training probes, typically logistic regression classifiers, to detect specific substructures within the model correlated with hallucinatory behavior. Semantic entropy probes (SEPs) can be used for such a task. Previous work using SEPs have found that they are generally predictive of hallucinatory behavior[7]; however, a detailed exploration of different mechanisms within models has yet to be accomplished. Therefore, I will apply probes across attention heads and weight subsets throughout the model. The probes that are the most predictive of semantic entropy will indicate which parts of the model are the most responsible. This will be done on several open source LLMs, including Llama-3[9] and Mistral 7B[10], to explore inherent variations across models. This work will require access to adequate GPU resources for running the models and training the classifier probes. I plan to apply only to graduate schools with access to these resources.

The expected outcome of **Aim 1** will be fine-grained understanding into which activations in the model are most responsible for predictions that can be categorized as hallucinations, across a spectrum of data. Thus, when these activations are high, we can notify the user of potential hallucinations and the possibility that the model is nearing its knowledge boundary. When the model's knowledge becomes "fuzzy", we can also cross compare with the input it is receiving from the handpicked dataset to interpret whether or not it's reaching its knowledge boundary. To this end, I will develop a model plugin that will output a hallucination metric for any prompt given, based on semantic entropy. This achieves more

granularity than previous binary methods. This information can then be leveraged to provide more transparency to a user regarding the model's internal knowledge base.

Aim 2: Discount substructures to mitigate hallucinations. I will leverage the results obtained from Aim 1 to develop a mitigation strategy whereby the activations of substructures most responsible for hallucinations are altered at inference time. Previous work has explored Inference-Time Intervention, whereby model activations are shifted in a truthful direction[11]. This research, however, was limited to the "truthful" perspective, with an emphasis on decoding inconsistency with true/false questions. By contrast, Aim 2 will generalize activation modification to a hallucination-correlated dimension across responsible model mechanisms, as discovered by Aim 1. I will do this by applying a discount factor to correlated attention heads and linear layers, such that these activations are either weakened or strengthened at inference time. By doing so, we encourage the model to be more certain, pushing it out of the boundary space. Evaluation of this method will first be performed with the Natural Questions subset described in Aim 1, in order to compare performance across the knowledge spectrum. Then I will evaluate with several popular hallucination-focused datasets, such as SQuADRUn[12] or TruthfulQA[13], to establish generalizability and show comparison to previous methods.

Intellectual Merit: ***The key contribution of this research is a novel methodology for detecting and mitigating hallucinatory behavior in language models.*** The detection of substructures relating to hallucinations will innovate on current mechanistic interpretability techniques. The method developed by **Aim 1** will provide critical insight into the knowledge space of state-of-the-art LLMs. This is useful in areas of model interpretation and in the development of model safety protocols. By providing users with meta information regarding the model's uncertainty level, this method can reduce the spread of misinformation by language models.

Aim 2 presents an approach that is 1) completely unsupervised, depending only on the semantic entropy probes and requiring no labels, and 2) agnostic to model or dataset. Furthermore, no previous work has investigated hallucination mitigation from the perspective of discounting responsible activations. To ensure success, I will work with experts in my future graduate research lab and seek collaborators from other labs.

References: [1] Spracklen, J., et al. (2024). *arXiv preprint arXiv:2406.10279.* [2] Huang, L., et al. (2023). *arXiv preprint arXiv:2311.05232.* [3] Shuster, K., et al. (2021). *EMNLP 2021.* [4] Pacchiardi, L., et al. (2024). *ICLR 2024.* [5] Farquhar, S., (2024). *Nature.* [6] Azaria, A. & Mitchell, T. (2023). *EMNLP 2023.* [7] Kossen, J., et al. (2024). *ICML 2024.* [8] Kwiatkowski, T., et al. (2019). *TACL 2019.* [9] Dubey, A., et al. (2024). *arXiv preprint arXiv:2407.21783.* [10] Jiang, A. Q., et al. (2023). *arXiv preprint arXiv:2310.06825.* [11] Li, K., et al. (2023). *NIPS 2023.* [12] Rajpurkar, P., et al. (2018). *ACL 2018.* [13] Lin, S., et al. (2022). *ACL 2022.*