Project Proposal: Cognitive Framework for Intrinsic Hallucination Detection in LLMs

Prateek Pandey
Roll Number: CS24M034

November 27, 2025

Abstract

As Large Language Models (LLMs) are increasingly deployed in critical applications, their tendency to generate "hallucinations"—plausible-sounding but factually incorrect assertions—poses a severe reliability challenge. Current detection methodologies face a "scalability wall," relying heavily on expensive, high-latency retrieval from external knowledge bases or opaque NLI models. In this proposal, we introduce a novel, intrinsic detection framework grounded in cognitive science. We propose to model "Attention Bias," simulating the hierarchical "skim and stare" reading behavior of humans. By integrating a Global Attention Bias (GAB) module for context selection and a Local Attention Bias (LAB) module for token-level saliency into a transformer architecture, we aim to achieve high-accuracy, interpretable hallucination detection without live external knowledge dependencies.

1 Introduction

Hallucination detection in text refers to the task of identifying and validating information that is inaccurately or falsely represented within textual content. While traditional approaches leverage explicit knowledge sources like Knowledge Graphs or Wikipedia, they face sustainability challenges due to the constant need for up-to-date knowledge.

We propose an alternative approach: verifying claims based on the intrinsic consistency of the text using cognitive signals. Our hypothesis is that human gaze patterns contain latent signals regarding textual inconsistency. When humans verify facts, they do not read uniformly; they exhibit specific attention biases, focusing intensely on entities, numbers, and relations that require verification. We aim to capture and model this behavior to guide a neural network.

2 Methodology

Our proposed architecture is a modular deep learning framework comprising three distinct phases: Data Acquisition, Global Attention Modeling, and Local Attention Modeling.

2.1 Phase 1: Cognitive Signal Acquisition (The Gaze Corpus)

To ground our models in human behavior, we will first construct a specialized dataset, the *Hallucination Gaze Corpus*.

- **Stimuli Selection:** We will sample claim-context pairs from the FactCC dataset and ensure diversity in topic and complexity to capture a broad range of hallucination types.
- **Data Collection Protocol:** We will employ high-frequency eye-tracking (e.g., SR Research Eyelink) to record annotators as they verify claims.
- **Target Metrics:** We will specifically extract *Fixation Duration* and *Fixation Count* at the token level. We anticipate that longer fixations will correlate with semantic inconsistencies.

2.2 Phase 2: Global Attention Bias (GAB) Model

The GAB model simulates the human "skimming" phase. Its objective is to identify the single most relevant sentence within a potentially long context paragraph, filtering out noise.

2.2.1 Ensemble Architecture

We will implement an ensemble of five state-of-the-art sentence transformers to compute cosine similarity between the claim (C) and every sentence in the context ($S_1, S_2, ...S_n$). The proposed models are:

1. `all-roberta-large-v1`
2. `all-mpnet-base-v1`
3. `gtr-t5-large`
4. `all-mpnet-base-v2`
5. `LaBSE` (Language-Agnostic BERT Sentence Embedding)

2.2.2 Voting Mechanism

We define a voting function $V(S_i)$ where a sentence S_i receives a vote if a model ranks it as the most similar to the claim. The sentence with the majority of votes is selected as the *Evidence Sentence* (S_{ev}):

$$S_{ev} = \text{argmax}_{S_i} \sum_{m=1}^{5} \mathbb{I}(\text{rank}_m(S_i) = 1)$$

This approach is robust against individual model biases and ensures that the downstream classifier focuses only on relevant information.

2.3 Phase 3: Local Attention Bias (LAB) Model

The LAB model simulates the "staring" phase, predicting token-level importance scores.

2.3.1 Training Strategy

We train a neural module to predict normalized fixation duration for every token.

- **Pre-training:** We will initialize the LAB module using general reading gaze datasets such as the PROVO Corpus and GECO Corpus.
- **Fine-tuning:** The module will then be fine-tuned on our Hallucination Gaze Corpus to adapt to cognitive patterns associated with fact-checking.

2.3.2 Feature Integration

Let $F_{lab} = \{f_1, f_2, ..., f_n\}$ be the predicted fixation scores. The LAB output modifies contextual embeddings E_{embed} by weighting them with F_{lab} before final classification.

3 End-to-End Classification Architecture

The final system integrates the components into a unified pipeline:

1. **Input:** A Claim (C) and a multi-sentence Context (Ctx).
2. **Step 1 (GAB):** The Ensemble Voting mechanism selects the most relevant sentence S_{ev}.
3. **Step 2 (Tokenization):** The input is formatted as $[CLS]\ C\ [SEP]\ S_{ev}\ [SEP]$.
4. **Step 3 (LAB):** The LAB module predicts saliency scores for these tokens.
5. **Step 4 (Classification):** A BERT-based Sentence Pair Classifier generates contextual embeddings. These are fused with the LAB scores and passed through a Feed-Forward Neural Network (FFNN) with Sigmoid activation:

$$Y = \text{Sigmoid}(W(F_{lab} \odot E_{embed}) + b)$$

4 Expected Results and Impact

4.1 Performance Targets

We aim to surpass the performance of existing baselines on the FactCC benchmark, specifically targeting a balanced accuracy exceeding 87%. We expect stronger performance in difficult cases such as entity swaps and number falsifications.

4.2 Interpretability

A key deliverable is explainability. Because the LAB module explicitly assigns weights to tokens, we can generate **Attention Heatmaps**. If the model flags a hallucination, the heatmap highlights conflicting words (e.g., mismatched dates), supporting user trust.

5 Conclusion

This proposal outlines a shift from "black-box" verification to a cognitively grounded framework. By modeling the "skim and stare" heuristic of human readers, we aim to build a hallucination detection system that is accurate, efficient, and interpretable, enhancing reliability in high-stakes domains.