Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Sztuczna Inteligencja (SZT)**
Specjalność: **-**

PRACA DYPLOMOWA MAGISTERSKA

Analiza metod agregacji map atencji do detekcji kontekstowych halucynacji w wielkich modelach językowych

Mikołaj Langner

Opiekun pracy
dr inż. Jan Kocoń

Słowa kluczowe: LLM, NLP, Detekcja halucynacji, Mapa atencji, Kontekstowe halucynacje

WROCŁAW 2025

Wrocław University of Science and Technology
Faculty of Information and Communication Technology

Field of study: **Artificial Intelligence**
Speciality: **-**

MASTER THESIS

Analysis of attention map aggregation methods for contextual hallucination detection in LLMs

Mikołaj Langner

Supervisor
dr inż. Jan Kocoń

Keywords: LLM, NLP, Hallucination detection, Attention map, Contextual hallucination

WROCŁAW 2025

Streszczenie

Wielkie modele językowe osiągają wysoką skuteczność w zadaniach przetwarzania języka naturalnego, lecz często generują halucynacje — odpowiedzi, które brzmią wiarygodnie, ale nie są poparte faktami lub dostarczonym kontekstem. W systemach typu Retrieval-Augmented Generation (RAG) problem ten jest szczególnie istotny, ponieważ model powinien opierać się wyłącznie na dostarczonej wiedzy.

Praca skupia się na wykrywaniu *halucynacji kontekstowych*, tj. sytuacji, w których odpowiedź modelu odbiega od dostarczonego kontekstu. Proponowane są cztery techniki agregacji uwagi: suma, podobieństwo kosinusowe, entropia oraz odległość Jensena-Shannona, które wykorzystują rozkłady atencji do identyfikacji nieprawidłowości w odpowiedzi modelu. Agregowane cechy są wykorzystywane w trybie nadzorowanym do detekcji halucynacji kontekstowych.

Eksperymenty obejmują różne modele językowe (Llama 2/3, Gemma, Phi, Bielik), dwa języki (angielski i polski) oraz dwa typy zadań (odpowiadanie na pytania i streszczanie). Badania uwzględniają generalizację między zadaniami i językami oraz odporność na długość kontekstu. Wyniki pokazują, że proste metody, takie jak Entropia i Suma, są konkurencyjne względem bardziej złożonych podejść i stanowią skuteczne, interpretowalne rozwiązanie do wykrywania halucynacji w systemach RAG.

Abstract

Large Language Models (LLMs) achieve state-of-the-art performance in many natural language processing (NLP) tasks, but are prone to hallucinations — confident, plausible outputs that are either false or not supported by the given context. In Retrieval-Augmented Generation (RAG) settings, this issue becomes particularly critical, as models are expected to generate answers grounded in explicitly retrieved passages.

This thesis focuses on detecting *contextual hallucinations*, where the generated responses deviate from the supporting context. This thesis proposes four attention-based aggregation techniques: sum, cosine similarity, entropy, and Jensen-Shannon distance, which leverage internal attention distributions to capture context misalignment. These signals are used as features in a supervised classification framework for hallucination detection.

Experiments are conducted on a range of LLMs (Llama 2/3, Gemma, Phi, Bielik), languages (English and Polish), and tasks (question answering and summarization). The methods are assessed for cross-task and cross-lingual generalization, as well as robustness to context length. The results show that simple approaches like Entropy and Sum consistently perform well, often matching or outperforming more complex techniques. The findings support attention aggregation as an interpretable and lightweight approach to hallucination detection in RAG systems.

Contents

1. Introduction

1.1. Motivation

Large Language Models (LLMs) have significantly advanced the field of natural language processing (NLP), enabling human-like text generation across a wide range of domains. These models are built on the Transformer architecture, which relies heavily on multi-head self-attention mechanisms to model complex dependencies between tokens in a sequence. Despite their impressive capabilities, LLMs remain prone to hallucinations, i.e., outputs that are syntactically correct but factually or contextually inaccurate. This phenomenon presents a major barrier to the safe and reliable deployment of LLMs in critical domains such as education, healthcare, and finance.

In Retrieval-Augmented Generation (RAG) frameworks, where LLMs are supplied with external documents or knowledge snippets at inference time, hallucinations take on a more specific form: contextual hallucinations. These arise when a model fails to ground its response in the provided context, producing outputs that contradict, distort, or ignore the relevant information despite having access to it. Unlike factual hallucinations that result from knowledge gaps, contextual hallucinations reflect a failure in internal reasoning or attention alignment.

Existing detection methods often rely on dynamic retrieval pipelines, external fact-checking tools, or log-likelihood estimation. However, these approaches are typically designed to flag general factual inconsistencies and may struggle to identify subtle mismatches between a model's response and the explicitly provided context. This highlights the need for more intrinsic detection strategies, especially those that can diagnose a model's failure to properly utilize accessible information.

Attention maps, which represent how LLMs distribute focus across tokens during inference, offer a promising yet underexplored pathway to hallucination detection in RAG settings. By aggregating and analyzing these attention patterns, it may be possible to uncover latent signals that correlate with hallucinated spans, potentially enabling more efficient and interpretable detection without reliance on external validators.

1.2. Problem Statement

In RAG settings, contextual hallucinations present a distinct challenge. Despite receiving a user query and a context retrieved from external sources, LLMs may still produce responses that are inconsistent with the evidence provided. Unlike hallucinations caused by missing or outdated knowledge, contextual hallucinations reflect a model's failure to adequately leverage and align its output with the input context, even when that context is explicitly available.

Although most existing detection methods operate at the output level, these methods often lack the granularity to capture subtle context-response mismatches. Consequently, there has been an increasing interest in using internal model signals to better understand and detect

hallucinations. Prior work has explored various signals, including attention distributions and hidden states, with the aim of tracking how a model processes and utilizes input information.

Attention maps, in particular, are attractive because of their interpretability and direct availability during inference. However, transformer models produce large attention matrices for each head and layer, and there is no established standard for how to aggregate or interpret these signals in the context of hallucination detection. The effectiveness of different aggregation strategies remains unclear, especially in RAG environments.

This thesis addresses this gap by systematically analyzing attention map aggregation methods for detecting contextual hallucinations in RAG scenarios. The goal is to determine whether specific aggregation strategies can highlight attention patterns that are predictive of hallucinated output and to evaluate how well attention-based indicators correlate with failures in grounding relative to the provided context.

1.3. Thesis Objectives

The central aim of this thesis is to investigate whether internal representations (i.e. attention maps) can be used to detect contextual hallucinations in LLMs operating within a RAG framework. In this setup, the model is guided by both a user query and retrieved context documents, which are intended to ground its responses in accurate and contextually appropriate information. However, models can still produce hallucinated outputs that, despite appearing plausible, fail to align with the provided context, reflecting a misalignment between the model's internal representations and the supplied evidence. These context-inconsistent outputs are the primary target of this investigation.

To achieve this aim, the thesis pursues the following specific objectives:

- **To define and formalize the concept of contextual hallucination in RAG settings.** This includes establishing a working definition of hallucination in relation to generation that diverges from the retrieved context, distinguishing it from more general hallucinations that arise from knowledge gaps or outdated training data.

- **To implement and systematically compare multiple attention map aggregation strategies.** As each head and layer of LLM generate a large attention map during inference, it is necessary to aggregate them. This thesis evaluates various techniques for aggregating these maps to extract meaningful patterns related to hallucination detection.

- **To design and execute a comprehensive experimental pipeline for evaluating aggregation strategies.** This includes testing how well the proposed aggregation techniques perform in detecting hallucinations relative to each other and to established baselines. The evaluation is carried out in multiple settings to examine whether some aggregation strategies consistently outperform others.

- **To assess the generalizability of attention-based aggregation methods across tasks and languages.** Experiments are carried out on both English and Polish datasets, and across distinct generation tasks, specifically: Question answering (QA)

and abstractive summarization. This objective aims to determine whether the observed performance of aggregation strategies is robust across linguistic and functional boundaries.

In summary, this thesis aims to bridge the gap between attention-based interpretability and hallucination detection in RAG systems by formalizing the concept of contextual hallucination, evaluating a range of aggregation techniques, and testing their effectiveness across multiple tasks and languages. In doing so, it seeks to provide both theoretical insights and practical tools for building more reliable, interpretable LLMs.

1.4. Scope

This thesis focuses on transformer-based decoder-only models applied to open-ended generation tasks. Hallucination is defined contextually, based on inconsistency with a prompt, document, or user query. The work does not explore hallucinations involving multimodal inputs (e.g., image captions) or multilingual outputs.

The analysis is limited to attention weights derived from inference-time outputs and does not involve model retraining or fine-tuning.

1.5. Contributions

This thesis builds on and significantly extends the work presented in the ICCS 2025 paper *"AggTruth: Contextual Hallucination Detection Using Aggregated Attention Scores in LLMs"* [33], in which I am a co-author. The key contributions, which span both the paper and the broader scope of this thesis, include:

- **Conceptualization of Contextual Hallucination in RAG.** Contributed to formalizing the concept of contextual hallucination within the specific context of RAG systems. This formalization provided the theoretical foundation for the *AggTruth* approach.
- **Development of Aggregated Attention Scoring Techniques.** Co-designed the core aggregation methods used in the *AggTruth* framework. These approaches were developed to effectively capture contextual grounding failures within attention distributions.
- **Implementation of a Comprehensive Experimental Pipeline.** Played a central role in implementing the experimental framework to evaluate attention aggregation methods and the construction of cross-task benchmarks. This thesis further extends this work to include both English and Polish datasets.
- **Empirical Evaluation and Cross-Domain Analysis.** Conducted extensive experiments to evaluate the generalizability of attention aggregation methods across different languages and tasks. This includes a detailed performance comparison between QA and summarization tasks, highlighting both the strengths and limitations of various aggregation strategies.

- **Open-Source Code and Reproducibility.** Contributed to the open source codebase accompanying the *AggTruth* paper, ensuring reproducibility and providing a foundation for future research in attention-based hallucination detection.

Together, these contributions represent a substantial step toward building more interpretable and reliable LLMs, particularly in the context of RAG systems, where grounding is critical to output fidelity.

1.6. Thesis Outline

This thesis is organized into seven chapters, each addressing a distinct aspect of the overall research question and the methodological approach.

1. **Introduction** provides an overview of the research context, describing the motivation, problem statement, objectives, and contributions of this thesis.

2. **Related Work** presents a comprehensive survey of prior work on hallucination detection in LLMs, focusing specifically on RAG systems. This chapter also covers related studies on attention mechanisms, hidden state analysis, and other internal signals relevant to grounding and interpretability.

3. **Foundations and Key Terms** formalizes the concepts of Transformer architecture, attention mechanism and contextual hallucination within the RAG frameworks, introducing the specific challenges of grounding responses in the provided context.

4. **Approaches and Methods** describes the technical foundations of the proposed attention aggregation methods. It also covers an overview of the examined baselines and other approaches.

5. **Experimental Setup** details the experimental setup, data sources and model architectures employed in this study. It presents the design and execution of empirical studies conducted for this thesis, including both task-specific (e.g. QA, summarization) and cross-linguistic (English and Polish) evaluations. It also discusses baseline comparisons.

6. **Results** provides a detailed analysis of the experimental outcomes, including quantitative performance metrics.

7. **Conclusions** synthesizes the findings of the previous chapters, reflecting on the broader implications of the results, limitations of the current approach, and potential avenues for future research. It summarizes the contributions of the thesis, revisits the main research questions, and offers final reflections on the impact of this work on the advancement of contextual hallucination detection in RAG systems.

Finally, the thesis concludes with a **Bibliography**, which provides a comprehensive list of the academic papers, datasets, and tools referenced throughout the document.

2. Related Work

2.1. Transformers and Attention

The introduction of the Transformer architecture by [36] marked a breakthrough in NLP by replacing sequential computation with parallelizable self-attention. This enabled significant performance improvements in tasks such as machine translation, summarization, and question answering. Unlike earlier models such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks [16], Transformers model token dependencies in a single forward pass, making them particularly suitable for long-context reasoning.

2.1.1. Self-Attention and Multi-Head Attention

At the core of Transformer models lies the *self-attention mechanism*, which allows each token to attend to every other token in a sequence. Each token is transformed into query, key, and value vectors, and their dot-product interactions define contextual relevance. Multi-head attention extends this idea by learning multiple independent attention patterns, capturing diverse linguistic signals.

Empirical analyses show that specific heads specialize in syntactic or semantic roles, such as resolution of coreference or positional encoding [8, 37]. However, other studies argue that many heads can be pruned with minimal performance loss [27], indicating redundancy in learned attention patterns.

2.1.2. Architectural Variants and Efficiency

To improve scalability, recent Transformer variants introduce optimizations such as sparse attention [7], flash attention [9], and low-rank approximations [39]. These advances support the continued growth of LLMs, enabling their use in real-time and resource-constrained settings.

2.2. Hallucinations in Large Language Models

Despite their strengths, LLMs often produce fluent but unfaithful outputs, commonly referred to as *hallucinations* [17, 19]. These outputs may contradict the input context or fabricate facts altogether. In RAG settings, such errors are especially problematic, as models are expected to ground their responses in explicit evidence.

Contextual hallucinations arise when LLMs fail to leverage the provided context, producing responses inconsistent with retrieved documents [32, 43]. These failures highlight deficiencies in context integration, attention focus, or grounding mechanisms within the model.

2.3. Hallucination Detection Techniques

Detecting hallucinations is inherently challenging due to language ambiguity, evaluation difficulty, and the opaque nature of LLM internal signals. In general, existing methods can be categorized as follows:

2.3.1. External Verification

These approaches compare generated outputs with external sources, such as structured knowledge bases or retrieved documents [22, 30]. Although effective, they are computationally expensive and often fragile in open-ended or loosely grounded tasks.

2.3.2. Internal Signal-Based Methods

Internal approaches use internal model signals to infer the likelihood of hallucination. These include attention distributions, hidden state patterns, and model-internal uncertainty scores.

Attention-Based Techniques

Attention weights have been used to gauge grounding: reliable generations tend to focus attention on source content. However, attention is not always faithful to the model's reasoning [18, 41]. Nevertheless, several techniques effectively leverage this signal:

- **Lookback Lens** [13]: Calculates the ratio of attention on prompt vs. generated tokens to assess grounding. Lightweight and interpretable, it works well for decoder-only models.

- **LapEigvals** [5]: Interprets attention maps as graphs and extracts spectral features (e.g., Laplacian eigenvalues) to capture structure. Strong performance but with added complexity.

- **Attention Rollout** [1]: Aggregates attention across layers to compute token-level influence. A foundational method for understanding and aggregating attention in Transformers.

Hidden-State-Based Methods

- **Truthfulness Detection from Activations** [4]: Trains classifiers on hidden layer activations to distinguish between truthful and hallucinated content.

- **Semantic Entropy and SEP** [20, 38]: Measures model uncertainty by clustering outputs (or approximating it via probes) to detect hallucination with high accuracy.

2.3.3. Fine-Tuning Approaches

Fine-tuning LLMs on labeled datasets can reduce hallucination [12, 25], but risks *catastrophic forgetting* [26]. Moreover, it is resource-intensive and may introduce dataset-specific biases. In contrast, this thesis emphasizes methods that use frozen LLMs and external classifiers trained on interpretable internal features.

2.4. Summary of Literature Review

The existing body of work highlights both the strengths and limitations of current approaches to Transformer architectures and hallucination detection in LLMs.

In terms of architecture, Transformer models have evolved significantly since their inception, with innovations such as multi-head self-attention, sparse approximations, and flash attention improving both performance and scalability. However, these advances come with increased model complexity and opacity, contributing to challenges in interpretability and control.

Hallucination detection methods can be grouped into three main categories: external verification, internal signal-based methods, and fine-tuning approaches. External methods leverage external data sources but are computationally intensive and brittle in open-domain scenarios. Internal methods offer more scalable alternatives by tapping into the model's own attention and activation signals. While attention-based techniques like Lookback Lens and Attention Rollout provide interpretability and efficiency, they are often used in limited settings or rely on simplified metrics. Hidden-state-based techniques, including entropy-based probes and activation classifiers, achieve high accuracy but typically demand model fine-tuning or extra supervision. Fine-tuning approaches have shown promise but pose risks such as catastrophic forgetting and poor generalizability.

Despite these developments, several challenges persist: limited evaluation beyond English and narrow domains, underutilized internal signals (especially attention), and trade-offs between detection accuracy and computational cost.

This thesis positions itself within this gap, aiming to develop interpretable, efficient, and multilingual attention-based methods for hallucination detection in retrieval-augmented generation (RAG) settings.

2.5. Research Problems

Although substantial progress has been made in hallucination detection, several key challenges remain:

- **Limited Generalizability**: Many methods are evaluated only in English or on narrow tasks.
- **Underexplored Internal Signals**: Attention-based signals are promising, but are often underutilized or oversimplified.
- **Efficiency vs. Accuracy Trade-offs**: Sampling-based techniques are accurate but slow; lightweight attention aggregations may offer better scalability.

2.5.1. Goals

This thesis addresses the identified limitations in the literature by developing interpretable and efficient attention-based features for hallucination detection in retrieval-augmented generation (RAG) systems. The goals are threefold:

1. Design novel aggregation strategies for attention maps. The first goal is to develop aggregation techniques that go beyond existing heuristics and better capture how attention relates to grounding in source context. By extracting higher-order features from attention distributions, the aim is to detect hallucinations more accurately while preserving model interpretability and runtime efficiency.

2. Evaluate across diverse tasks and languages. A major gap in the current literature is the limited evaluation of hallucination detection methods across multiple languages and task settings. This thesis addresses this by testing the proposed features on both English and Polish datasets, and across two distinct generation tasks: question answering and summarization. This contributes to a better understanding of the cross-linguistic and cross-task generalizability of attention-based signals.

3. Compare with strong baselines and leverage Semantic Entropy. To benchmark performance, the proposed approach is compared with established methods such as Lookback Lens. Additionally, Semantic Entropy is used as a supervisory signal for weakly supervised classifier training, allowing for robust performance without extensive labeled data. This integration provides a principled way to combine interpretability with statistical grounding.

3. Foundations and Key Terms

Accurate detection of contextual hallucinations in RAG systems requires a detailed understanding of the internal mechanisms by which LLMs integrate contextual information and generate outputs. This chapter presents the foundational components underpinning such systems, including the Transformer architecture, self- and multi-head attention mechanisms, and the formal notion of contextual hallucination. These concepts serve as the basis for the attention aggregation techniques proposed in the next chapter.

3.1. Core Concepts

3.1.1. Transformer Architecture

The Transformer architecture [36] underlies most contemporary LLMs, such as BERT [11] and GPT [6]. Its central innovation, the self-attention mechanism, enables efficient modeling of long-range dependencies without recurrence.

Each layer in a Transformer comprises:

- A multi-head self-attention sublayer,
- A position-wise feed-forward network (FFN),

each followed by residual connections and layer normalization.

To encode token order, positional embeddings are added to the input token representations. These may be fixed sinusoidal functions or learned embeddings.

Sinusoidal Positional Encoding [36] is computed as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{3.1}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{3.2}$$

Although later models such as RoBERTa [24] and Llama [34, 35] adopt learned positional embeddings, the sinusoidal variant remains widely studied.

Feed-Forward Networks (FFNs) apply nonlinear transformations independently to each position:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{3.3}$$

where $W_1 \in \mathbb{R}^{d\times d'}$, $W_2 \in \mathbb{R}^{d'\times d}$, $b_1 \in \mathbb{R}^{d'}$, $b_2 \in \mathbb{R}^d$ are learnable parameters.

Layer Composition: Let $\mathbf{H} \in \mathbb{R}^{T\times d}$ denote the token representations. A single Transformer layer applies:

$$\mathbf{H}' = \text{LayerNorm}(\mathbf{H} + \text{MultiHead}(\mathbf{H})) \tag{3.4}$$

$$\mathbf{H}'' = \text{LayerNorm}(\mathbf{H}' + \text{FFN}(\mathbf{H}')) \tag{3.5}$$

In decoder-only models (e.g., GPT), attention is constrained via causal masking to ensure that predictions do not depend on future tokens.

Figure 3.1 illustrates the overall data flow in a Transformer model. A detailed view of the internal operations within each layer is shown in Figure 3.2.

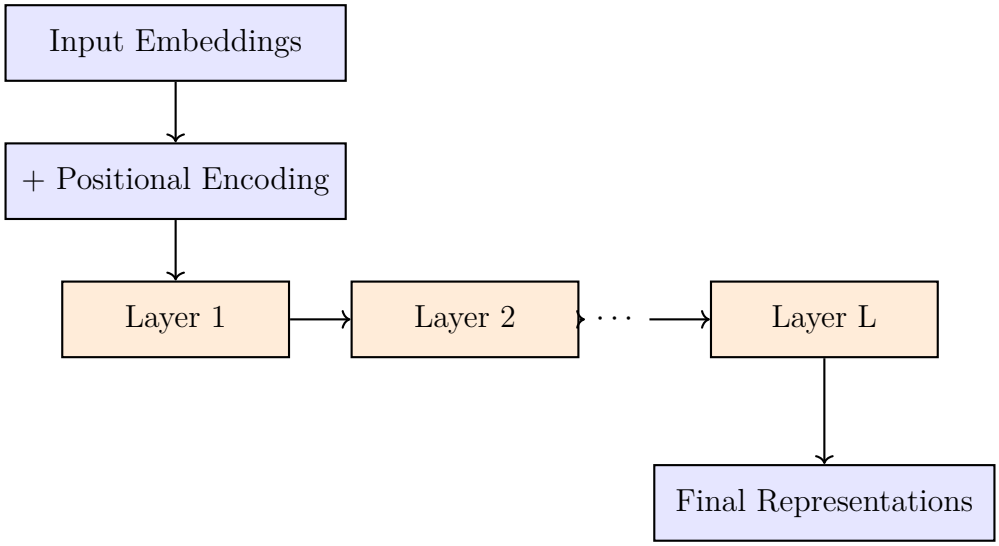

Figure 3.1: Transformer architecture. Input embeddings flow through a stack of L layers, each with attention and FFN sublayers.

3.1.2. Attention Mechanism

Self-attention enables a model to compute contextualized representations by weighting other tokens in the sequence based on their relevance. For a sequence represented by $\mathbf{H} \in \mathbb{R}^{T \times d}$, query, key, and value matrices are computed as:

$$Q = \mathbf{H}W^Q, \quad K = \mathbf{H}W^K, \quad V = \mathbf{H}W^V$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ are trainable parameters.

The scaled dot-product attention is then defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3.6}$$

This mechanism, illustrated in Figure 3.3, produces a weighted sum of the values, where the weights are derived from the similarity between queries and keys.

Attention matrices offer insight into the internal decision-making processes of the model, providing a basis for interpretability and, in this thesis, for the detection of hallucinations.

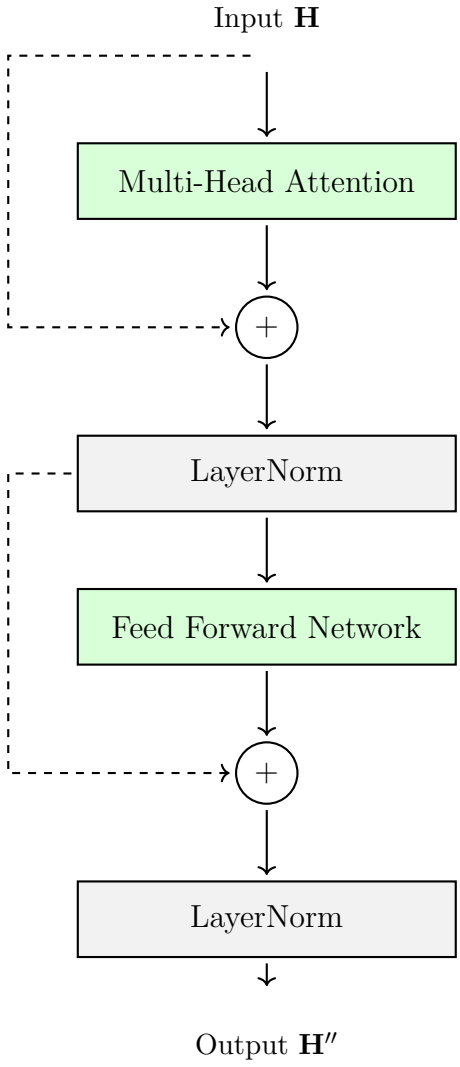

Figure 3.2: Internal composition of a Transformer layer with residual connections (dashed lines) and layer normalization.

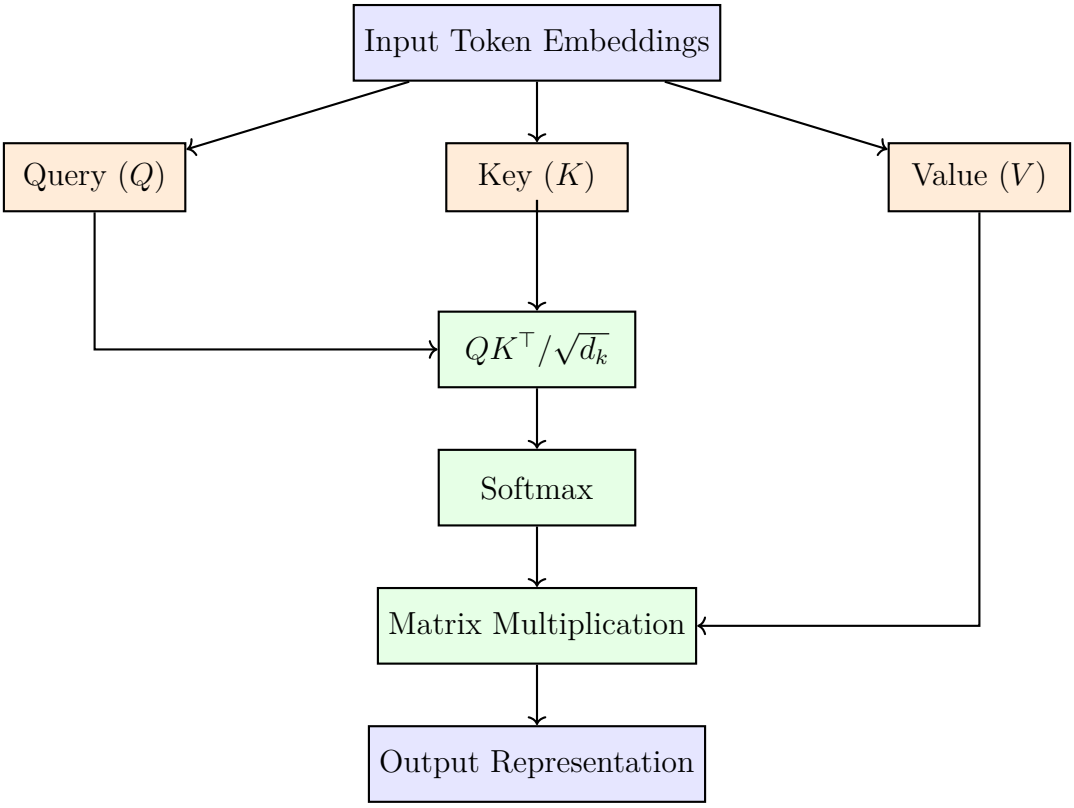

Figure 3.3: Scaled dot-product attention mechanism: Queries and keys determine attention weights, which are used to compute a weighted sum over values.

3.1.3. Multi-Head Attention

Rather than computing a single attention distribution, Transformers employ h parallel attention heads:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.7}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{3.8}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$ and the output projection $W^O \in \mathbb{R}^{hd_k \times d}$ integrates these subspace-specific representations.

Empirical studies (e.g., [8, 37]) have shown that different attention heads specialize in distinct linguistic patterns, a property that this thesis leverages through the attention aggregation techniques.

3.1.4. Retrieval-Augmented Generation (RAG)

RAG systems [22, 32] enhance generation by retrieving contextually relevant external documents. Given an input query q, the system retrieves a set of documents $C = \{c_1, ..., c_n\}$, which are then used to condition the response r.

- **Retriever:** Identifies and returns documents C relevant to the query q.
- **Generator:** Conditions on both q and C to generate a response r.

Figure 3.4 shows the overall structure of a RAG pipeline.

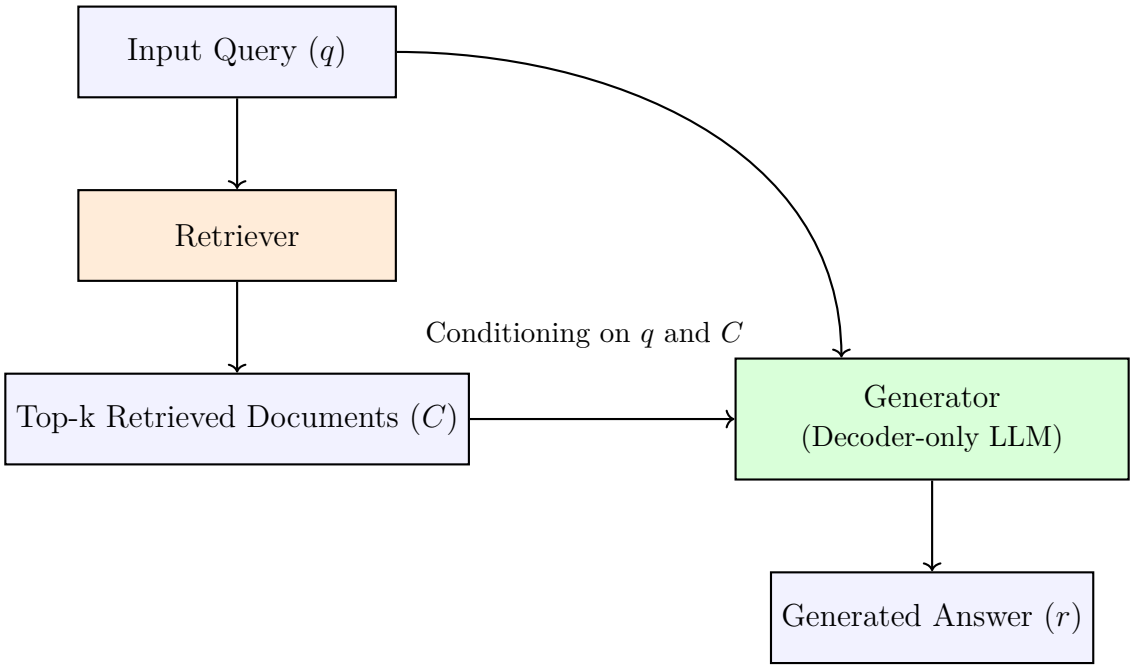

Figure 3.4: Structure of a RAG system. The retriever identifies relevant documents, which are used as grounding context by the generator.

3.1.5. Contextual Hallucinations

Contextual hallucination occurs when a generated response r is fluent but inconsistent with the combination of the query and retrieved context. Formally:

$$r \not\models (q \wedge C)$$

This indicates that the generated content introduces facts not entailed by the query or retrieval, regardless of linguistic coherence.

In this thesis, hallucination detection is cast as a binary classification problem that uses internal model signals to assess whether a response is contextually grounded.

3.2. Mathematical Formulation

Let q denote the input query, C the retrieved context, and r the generated response. Define a hallucination label as a binary function:

$$h(r, q, C) = \begin{cases} 1 & \text{if } r \not\models (q \wedge C) \\ 0 & \text{otherwise} \end{cases}$$

We aim to learn an approximation of h using a model f_θ, which operates on a feature representation $\phi(r, q, C)$:

$$h(r, q, C) \approx f_\theta(\phi(r, q, C))$$

The feature function ϕ maps internal model signals, such as attention weights, to an interpretable representation. The attention aggregation techniques proposed in the following chapters define novel instances of ϕ, which are used to train classifiers for hallucination detection.

4. Approaches and Methods

This chapter presents the methods developed to detect contextual hallucinations in RAG systems. The core contribution is a set of novel attention-based aggregation techniques designed to exploit structural signals in multi-head attention layers of decoder-only Transformer models. These techniques —*Sum*, *Cosine Similarity*, *Entropy*, and *Jensen–Shannon Distance*—are intended to reduce the dimensionality of the attention maps while retaining alignment-relevant information.

To contextualize these methods, we compare them with a recent state-of-the-art approach based on internal signals from the literature — *Lookback Lens* [13].

The chapter concludes with a description of the supervised learning framework used to predict hallucination labels at the token level, along with a modular pipeline that combines inference, attention extraction, feature aggregation, and classification.

4.1. Definitions and Notation

Let a response generated by a decoder-only Transformer contain N tokens. For each token, attention is distributed on C context tokens retrieved by a RAG system. The resulting attention data can be represented as a tensor of shape $L \times H \times N \times C$, where L is the number of layers and H the number of attention heads.

Let $a_{l,h,t,i}$ denote the attention weight assigned by layer l, head h, and output token t to the i-th context token. The vector $\mathbf{a}_{l,h,t} \in \mathbb{R}^C$ represents the attention distribution of the head h in layer l for the token t.

Due to the scale of this tensor and the presence of redundant or noisy dimensions, we aggregate attention selectively to produce compact and informative feature vectors for hallucination prediction. The methods used for this aggregation are described in section 4.3, with baseline comparisons in section 4.2 and the complete detection pipeline in section 4.5.

4.2. Baseline Methods

We benchmark our proposed techniques against a method based on internal signals from recent work: *Lookback Lens* [13]. This baseline relies on the Transformer attention behavior but differs in its assumptions and objectives.

4.2.1. Lookback Lens

Lookback Lens identifies hallucinations based on the intuition that grounded responses should pay more attention to context passages than to previously generated tokens. The method computes a *lookback ratio* to quantify this relative attention.

Let the model input include the system prompt, the retrieved context $\mathcal{X}$, and previous generations $\mathcal{Y}_{<t}$. For head h in layer l in generation step t, average attention is calculated as:

$$A_t^{(l,h)}(\text{context}) = \frac{1}{N}\sum_{i=1}^{N}\alpha_{t,i}^{(l,h)}, \quad A_t^{(l,h)}(\text{new}) = \frac{1}{t-1}\sum_{j=N+1}^{N+t-1}\alpha_{t,j}^{(l,h)}$$

Then the lookback ratio is:

$$\text{LR}_t^{(l,h)} = \frac{A_t^{(l,h)}(\text{context})}{A_t^{(l,h)}(\text{context}) + A_t^{(l,h)}(\text{new})}$$

These values are collected in a matrix $\mathbf{v}_t \in \mathbb{R}^{L\times H}$ and averaged over spans to generate feature vectors, which are classified by logistic regression:

$$P(y = 1 \mid \bar{\mathbf{v}}) = \sigma(\mathbf{w}^\top \bar{\mathbf{v}} + b)$$

.

Compared to the approach of this thesis (see section 4.3), the Lookback Lens computes attention over the entire input (prompt + context + generated tokens), whereas proposed methods isolate attention to context tokens to ensure a more grounded signal.

4.2.2. Semantic Entropy

Semantic Entropy (SE) [38] estimates the uncertainty of the model output at the semantic level, not just syntactic or lexical. It operates by clustering sampled completions based on their entailment relationships and computing an entropy score over the cluster distribution:

$$\mathcal{H}_{SE}(x) = -\sum_{k=1}^{K} p(C_k|x)\log p(C_k|x)$$

where C_k are semantic clusters formed via entailment checks.

This method is computationally expensive, so recent work introduced *Semantic Entropy Probes (SEPs)* [20]—linear models trained to predict SE from internal representations:

$$\text{SEP}(\mathbf{h}_i) \approx \mathcal{H}_{SE}(x)$$

These models allow for semantic uncertainty estimation with a single forward pass. For binary classification, SE can be thresholded using an optimized value γ^*, chosen to minimize the classification variance on the training set.

4.3. Proposed Aggregation Techniques

In this thesis, four attention aggregation methods are proposed. They are designed to summarize passage-focused attention for each generated token. These techniques transform the attention vector $\mathbf{a}_{l,h,t} \in \mathbb{R}^C$ into scalar features interpretable by downstream classifiers.

4.3.1. Attention Scope

Unlike Lookback Lens, this work restricts attention analysis to context passages only. As shown in Figure 4.1, the submatrix of attention is isolated from generated tokens to retrieved documents, excluding the prompt and query components.

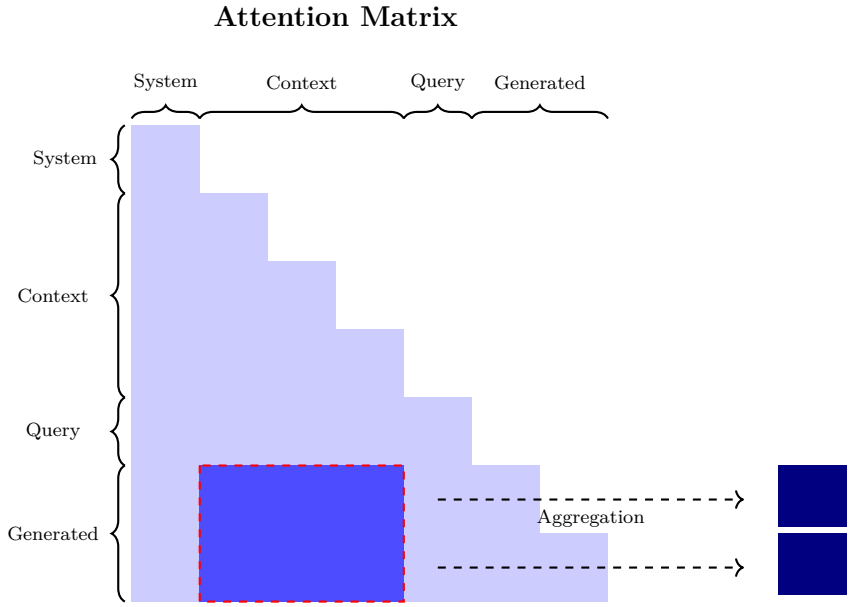

Figure 4.1: Attention from generated tokens to passage tokens (highlighted box) is selected and aggregated. Braces denote token type boundaries in the input and output sequences.

4.3.2. Sum

A simple method that measures total attention allocated to context tokens:

$$\text{Sum}_{l,h,t} = \sum_{i=1}^{C} a_{l,h,t,i}$$

This feature captures the degree of grounding: more attention to context suggests a greater reliance on retrieved evidence.

4.3.3. Cosine Similarity

This technique measures alignment between attention heads within the same layer:

$$\text{CosSim}_{l,h,t} = \frac{1}{H-1} \sum_{h' \neq h} \frac{\mathbf{a}_{l,h,t} \cdot \mathbf{a}_{l,h',t}}{\|\mathbf{a}_{l,h,t}\| \|\mathbf{a}_{l,h',t}\|}$$

The assumption is that divergence in attention across heads may indicate ambiguity or misalignment.

4.3.4. Entropy

To assess distributional uncertainty, entropy is calculated over an extended attention vector that includes a *missing mass* term:

$$\tilde{\mathbf{a}}_{l,h,t} = \left[a_{l,h,t,1}, \ldots, a_{l,h,t,C}, 1 - \sum_{i=1}^{C} a_{l,h,t,i} \right]$$

$$\text{Entropy}_{l,h,t} = -\sum_{i=1}^{C+1} \tilde{a}_{l,h,t,i} \log_2 \tilde{a}_{l,h,t,i}$$

Higher entropy indicates uncertainty or lack of focus.

4.3.5. Jensen–Shannon Distance

To detect a divergence from the consensus between heads, Jensen-Shannon distance is calculated between each head and the average of the entire layer:

$$a_{l,\text{ref},t,i} = \frac{1}{H} \sum_{h=1}^{H} a_{l,h,t,i}$$

$$\text{JS-Dist}_{l,h,t} = \sqrt{\frac{1}{2} \sum_{i=1}^{C} \left(a_{l,h,t,i} \log \frac{a_{l,h,t,i}}{m_i} + a_{l,\text{ref},t,i} \log \frac{a_{l,\text{ref},t,i}}{m_i} \right)}$$

where $m_i = \frac{1}{2}(a_{l,h,t,i} + a_{l,\text{ref},t,i})$.

High values may indicate abnormal behavior of specific heads.

4.4. Supervised Hallucination Detection

4.4.1. Formulation

Each token generated is represented by an attention-based feature vector $\mathbf{X} \in \mathbb{R}^{HL}$. The task is to classify whether it is hallucinated:

$$P(y = 1 \mid \mathbf{X}) = \sigma(\mathbf{w}^\top \mathbf{X} + b)$$

The model is trained using standard binary cross-entropy loss, with class weighting applied to mitigate label imbalance. A regularization strength of $C = 0.01$ is used to prevent overfitting (as suggested in [13]).

4.4.2. Design Rationale

Logistic regression is chosen for its interpretability and low computational overhead. The learned weights indicate which attention features correlate the most with the risk of hallucination.

4.5. End-to-End Detection Pipeline

This section outlines the end-to-end pipeline used to operationalize hallucination detection based on attention-derived features. The pipeline transforms an input query and retrieved context into aggregated features aligned with binary hallucination labels, which are then used to train a supervised classifier.

An overview of the system is provided in Figure 4.2, including its five main components: (1) prompt construction and generation, (2) external evaluation, (3) token-level alignment, (4) feature extraction, and (5) classification.

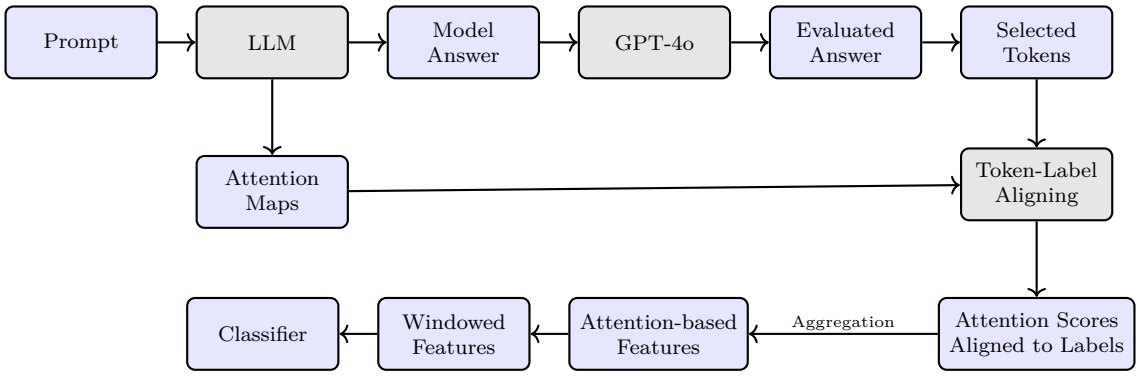

Figure 4.2: Overview of the hallucination detection pipeline from prompt construction to classifier-ready features. The blue boxes correspond to data objects whereas gray boxes correspond to particular processes.

Prompt Construction and Generation

Each instance begins with a user query accompanied by a set of retrieved context passages. This prompt is passed into a decoder-only language model, configured to expose internal attention weights during autoregressive generation. The model produces a free-form response token by token, logging full multi-head, multi-layer attention maps between generated tokens and the input sequence.

Response Evaluation

To generate training labels without human annotation, this work adopts the *LLM-as-a-Judge* paradigm. Specifically, GPT-4o is used as an external evaluator to assess whether the generated response is faithful to the retrieved context.

This approach is motivated by recent findings showing that GPT-4 reaches high agreement with expert human annotators when evaluating factual consistency, coherence, and relevance in text generation tasks. In [23], the authors report on the agreement over 85% between GPT-4 and expert judges, validating its role as a reliable automated evaluator. GPT-4o is prompted with the original query, the retrieved documents, and the response of the model, and returns a hallucination judgment at the span level.

Token-Level Label Alignment

Based on the GPT-4o judgment, each generated token is assigned a binary hallucination label. If a token occurs within a span marked as hallucinated, it receives a label of 1; otherwise, it is labeled 0. These token-level annotations are aligned with the attention weights of each head and layer, producing a four-dimensional tensor of shape $L \times H \times N \times C$, where L is the number of layers, H the number of heads, N the number of generated tokens, and C the number of context tokens.

Sliding Window Aggregation

To convert token-level attention tensors into usable classifier input, a sliding window is applied across the generated sequence. A window size of 8 tokens with a stride of 1 is used, consistent with prior work [13]. Each window is labeled as hallucinated if it contains at least one hallucinated token.

For each token in a window, the attention is aggregated using the techniques described in section 4.3. These values are averaged across the window to form a single feature vector per attention head, producing a fixed-size representation of attention behavior during generation.

Feature Preparation and Classification

Once window-level attention features are extracted, they are concatenated across layers and heads to form a final feature vector for classification. These vectors, aligned with their binary hallucination labels, are used to train a logistic regression model as described in section 4.4.

This framework is modular and lightweight: it requires no architectural modifications to the language model and leverages attention scores already computed during inference. The resulting classifier operates efficiently and interpretably, enabling integration into downstream RAG pipelines for the detection of hallucinations in real time.

5. Experimental Setup

This chapter details the experimental framework used to evaluate the proposed attention aggregation methods and baseline techniques for hallucination detection. The experiments are designed to test the generalizability and effectiveness of these methods across multiple language models, datasets, and task types, with an emphasis on cross-task and cross-lingual robustness.

The proposed approach to hallucination detection is based on extracting interpretable features from attention distributions in LLMs. Specifically, the methods involve novel aggregation techniques that quantify the grounding of generated tokens in retrieved context documents. These features are used to train lightweight external classifiers that do not require fine-tuning the base language model. To benchmark performance, the proposed techniques are compared with established methods such as Lookback Lens, and complemented with Semantic Entropy as a weak supervisory signal. The entire pipeline is implemented using Python-based tools, including PyTorch, HuggingFace Transformers, and Scikit-learn, enabling reproducible experiments across models, languages, and tasks. This setup allows for a systematic evaluation of how internal attention patterns can serve as reliable indicators of hallucination in multilingual RAG scenarios.

The chapter proceeds as follows: First, the datasets used (English and Polish) are described, and then the language models evaluated are introduced. Then, descriptions of the two task types (question answering and summarization) and the semantic entropy generation procedure are provided. Eventually, the evaluation metrics used are presented.

5.1. Datasets

To benchmark attention-based hallucination detection, six datasets were selected, covering English and Polish, and covering both QA and summarization tasks. This diversity ensures coverage of extractive and abstractive generation, as well as morphologically rich low-resource languages.

Table 5.1: Datasets overview and usage scope.

Dataset	**Language**	**Task Type**	**# Examples**
Natural Questions (NQ)	English	QA	1,819
CNN / Daily Mail (CNN/DM)	English	Summarization	1,000
HotPotQA	English	QA	1,050
XSum	English	Summarization	1,000
PolQA	Polish	QA	710
Polish Summary Corpus (PSC)	Polish	Summarization	569

NQ and CNN/DM are used primarily for training, while HotPotQA, XSum, PolQA, and PSC serve for evaluation under transfer and generalization scenarios.

5.2. Language Models

The experiments include five decoder-only Transformer models, chosen to reflect diversity in architecture size, multilingual capability, and real-world deployment relevance.

5.2.1. Llama

Llama-2-7B-chat [35] is used exclusively in the first experiment for compatibility with Lookback Lens implementation. Attempts to adapt the Lookback Lens baseline to other LLMs were hindered by model-specific architectural assumptions.

Llama-3.1-8B-Instruct [2] is a newer Meta model trained with improved multilingual and instruction-following capabilities. It is used in the rest of the experiments.

5.2.2. Phi

Phi-3.5-mini-Instruct [31], developed by Microsoft, is a compact and instruction-tuned model optimized for efficiency and zero-shot generalization.

5.2.3. Gemma

Gemma-2-9B-IT [10], released by Google DeepMind, is optimized for long-context comprehension and factual grounding.

5.2.4. Bielik

Bielik-11B-v2.3-Instruct [3] is a Polish-specific model based on the Mistral architecture, designed for low-resource language modeling.

5.3. Tasks

The detection methods are evaluated on two Natural Language Generation (NLG) tasks with high hallucination risk: question answering and abstractive summarization. Experiments include both same-task and cross-task transfer settings.

5.3.1. Question Answering

QA requires models to generate factually correct answers grounded in the retrieved context. The datasets used for this task are:

- **Natural Questions (NQ)** [21]: long-form QA pairs with Wikipedia context.

- **HotPotQA** [42]: Multi-hop QA requiring reasoning across documents.
- **PolQA** [40]: Polish-language QA dataset; used for cross-lingual evaluation.

5.3.2. Abstractive Summarization

Summarization involves generating a concise output that captures the meaning of a longer input. Hallucination is common in this task due to abstraction.

- **CNN / Daily Mail (CNN/DM)** [15]: news articles with summaries.
- **XSum** [28]: highly abstractive single-sentence summaries.
- **Polish Summary Corpus (PSC)** [29]: a Polish-language summarization benchmark.

5.4. Generating Semantic Entropy Labels

To support weak supervision experiments (see section 6.5), semantic entropy labels are generated, following the methodology of [20, 38].

For each prompt, $N = 10$ diverse responses are sampled, using temperature-based decoding (where $T = 1.0$). These responses are clustered using the DeBERTa-v2-XXL-MNLI model [14], which checks for entailment between each pair of responses in either direction. Responses with such entailment are grouped into semantic clusters $C_1, \ldots, C_K$.

The entropy over these clusters is calculated as:

$$\mathcal{H}_{SE}(x) = -\sum_{k=1}^{K} p(C_k|x) \log p(C_k|x)$$

where $p(C_k|x)$ is the proportion of completions in cluster C_k. These scores are binarized to produce hallucination labels in classification settings.

5.5. Evaluation Metrics

5.5.1. Area Under the ROC Curve (AUC)

The primary evaluation metric is the Area Under the Receiver Operating Characteristic Curve (AUC), which captures how well the model distinguishes hallucinated from grounded outputs across thresholds:

$$\text{AUC} = \int_0^1 TPR(FPR^{-1}(x))\, dx$$

Where:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

Justification

AUC is used because:

- It is threshold-independent.
- It is robust to class imbalance.
- Provides a meaningful probabilistic interpretation.

5.6. Experimental Conditions and Design

Experiments are organized to test detection robustness across:

1. **Context length sensitivity**
2. **Baseline comparison**
3. **Cross-task and cross-lingual transfer**
4. **Alternative supervision via semantic entropy**

5.6.1. Context Length Analysis

This work analyzes context length distributions across models and datasets, stratified by hallucination label. This reveals how retrieval size affects hallucination frequency and model behavior.

Additionally, it was evaluated how aggregation features degrade or improve with increasing context length. Stratification by length quartiles is used for:

- NQ → HotPotQA
- CNN/DM → XSum

5.6.2. Baseline Comparison: Lookback Lens

Using LLaMA-2-7B-chat, the proposed aggregation methods are compared with the Lookback Lens [13]. Due to architecture-specific constraints, this baseline is run only on Llama-2.

Train-test transfers evaluated:

- NQ → CNN/DM, HotPotQA, XSum
- CNN/DM → NQ, HotPotQA, XSum

5.6.3. Cross-Model and Cross-Task Generalization

The Llama-3, Phi, Gemma, and Bielik models were tested in three transfer regimes:

- **Same-Task:** NQ → HotPotQA, CNN/DM → XSum
- **Cross-Task:** NQ → CNN/DM, CNN/DM → NQ
- **Cross-Lingual:** NQ → PolQA, CNN/DM → PSC

5.6.4. Supervision from Semantic Entropy

In this experiment, classifiers are trained on semantic entropy scores instead of binary labels. Predictions are made on the standard binary hallucination task.

Two probe positions were evaluated:

- **Before-Generation Token (BFT)**
- **Second-to-Last Token (SLT)**

Sliding windows are not used here; each input is represented by a single token-level feature vector.

5.6.5. Hallucination Rate Distribution

Table 5.2 reports hallucination rates in all model-dataset pairs, using GPT-4o as an LLM-as-a-Judge evaluator (see section 4.5). PSC exhibits the highest error rate, underscoring the difficulty of Polish summarization.

Table 5.2: Hallucination rates across models and datasets (as judged by GPT-4o).

Dataset	Bielik	Gemma	Llama	Phi
CNN/DM	0.145	0.074	0.131	0.242
HotPotQA	0.153	0.147	0.437	0.257
NQ	0.206	0.156	0.241	0.390
PSC	0.876	0.808	0.407	0.329
PolQA	0.163	0.196	0.342	0.554
XSum	0.291	0.245	0.230	0.334

6. Results

This chapter presents the empirical evaluation of the proposed attention aggregation methods compared to baseline approaches, using the experimental framework outlined in chapter 5.

All experiments use the Area Under the ROC Curve (AUC) as the primary metric. Each section presents figures or tables followed by interpretive discussion. Patterns are analyzed by task and model, transferability, and the effect of supervision quality.

6.1. Context Length Distributions

Figure 6.1 shows the distribution of input context lengths for both hallucinated and non-hallucinated instances, stratified by dataset. The goal of this analysis is to determine whether hallucination correlates with the amount of contextual information available to the model at inference time.

Across datasets, the context length distributions for hallucinated and non-hallucinated samples are broadly similar. In particular, there is no clear shift or skew that would suggest a strong dependency between input size and hallucination likelihood. This indicates that hallucinations are not driven by extremes of evidence quantity—neither excessive nor insufficient context appears to be a consistent trigger.

This finding is important because it reinforces the assumption that hallucination is not a simple function of context length, and thus cannot be mitigated merely by truncating or extending retrieved documents. It justifies the need for more nuanced feature extraction methods, such as attention-based aggregation, that aim to model how the content of context contributes to hallucination.

6.2. Context Length Sensitivity

This section investigates whether the performance of the attention aggregation features varies with the amount of context provided to the language model. Specifically, the cumulative AUC scores across bins of increasing input context length for both tasks were examined. Figure 6.2 and Figure 6.3 present these results.

For the QA task, most aggregation methods exhibit stable performance across context lengths, suggesting that their signal quality is not tightly coupled with the volume of evidence. This is particularly evident for the Sum and Entropy features, which maintain consistent AUC scores regardless of context length.

In contrast, the summarization task shows a more nuanced interaction with context size:

- **Gemma** benefits from short-context scenarios where JS-Dist performs strongly, indicating that when evidence is concise, token-level distributional differences are more informative.

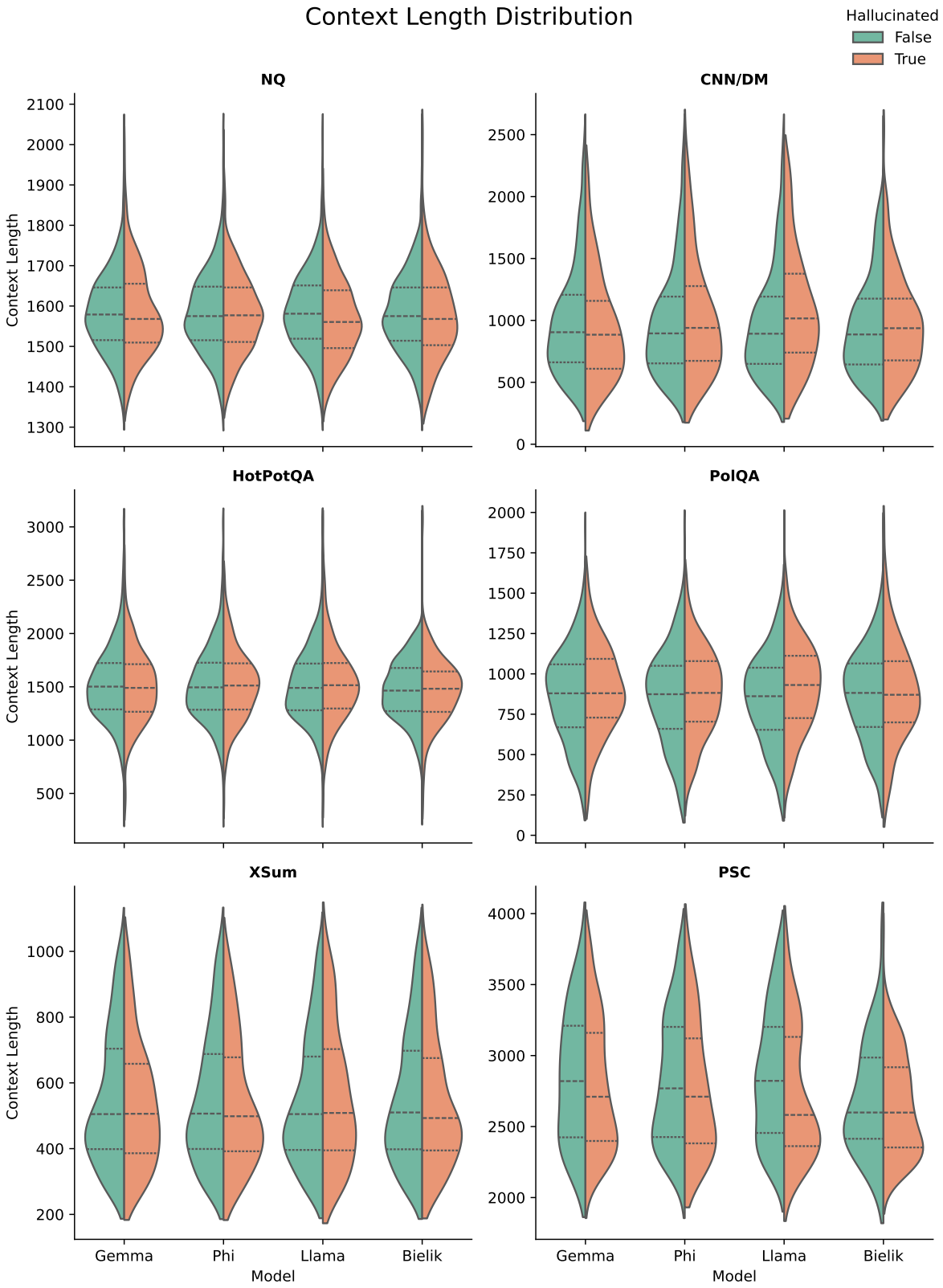

Figure 6.1: Distributions of context lengths stratified by dataset and hallucination label.

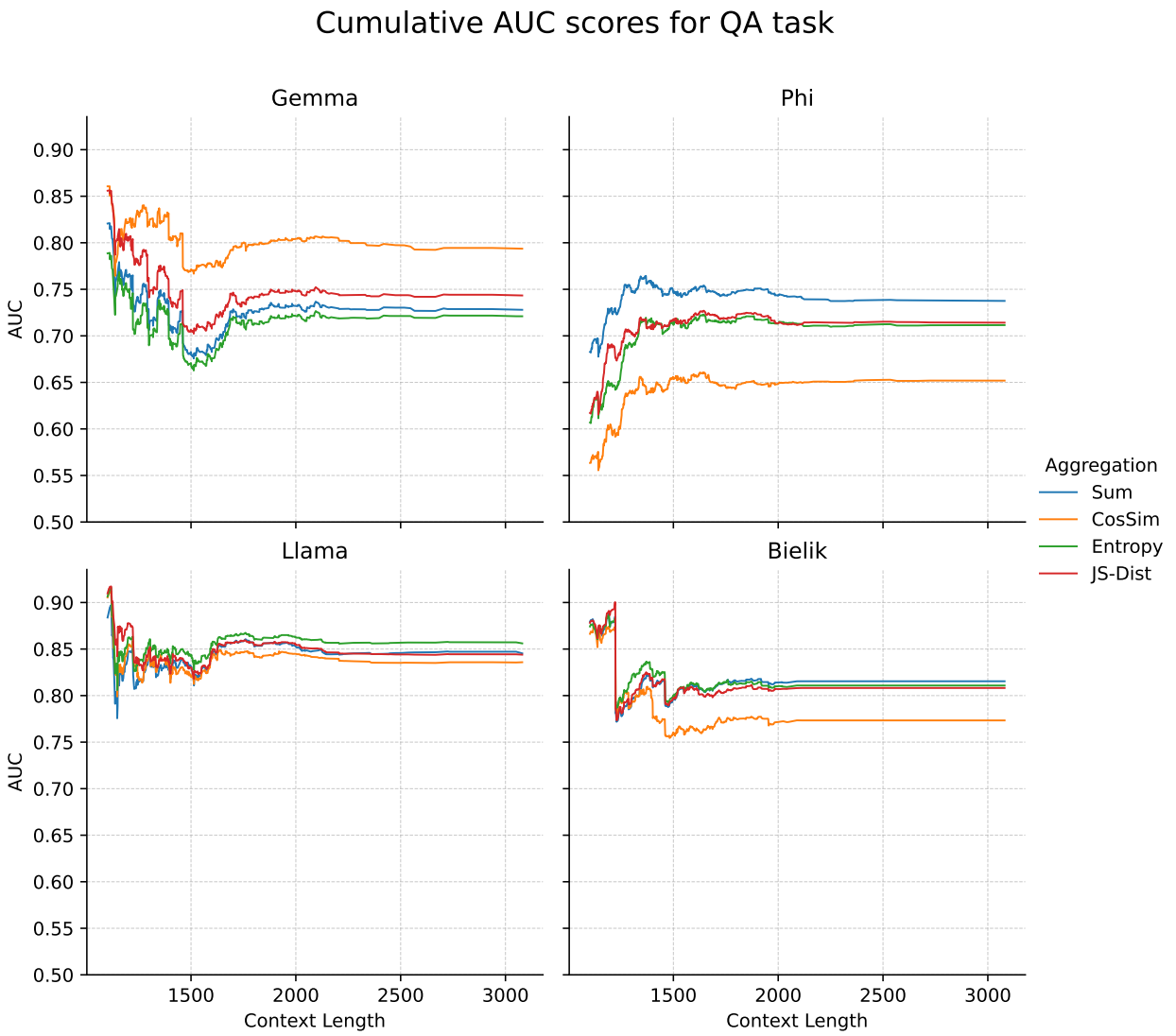

Figure 6.2: Cumulative AUC scores across context length for QA task (NQ → HotPotQA).

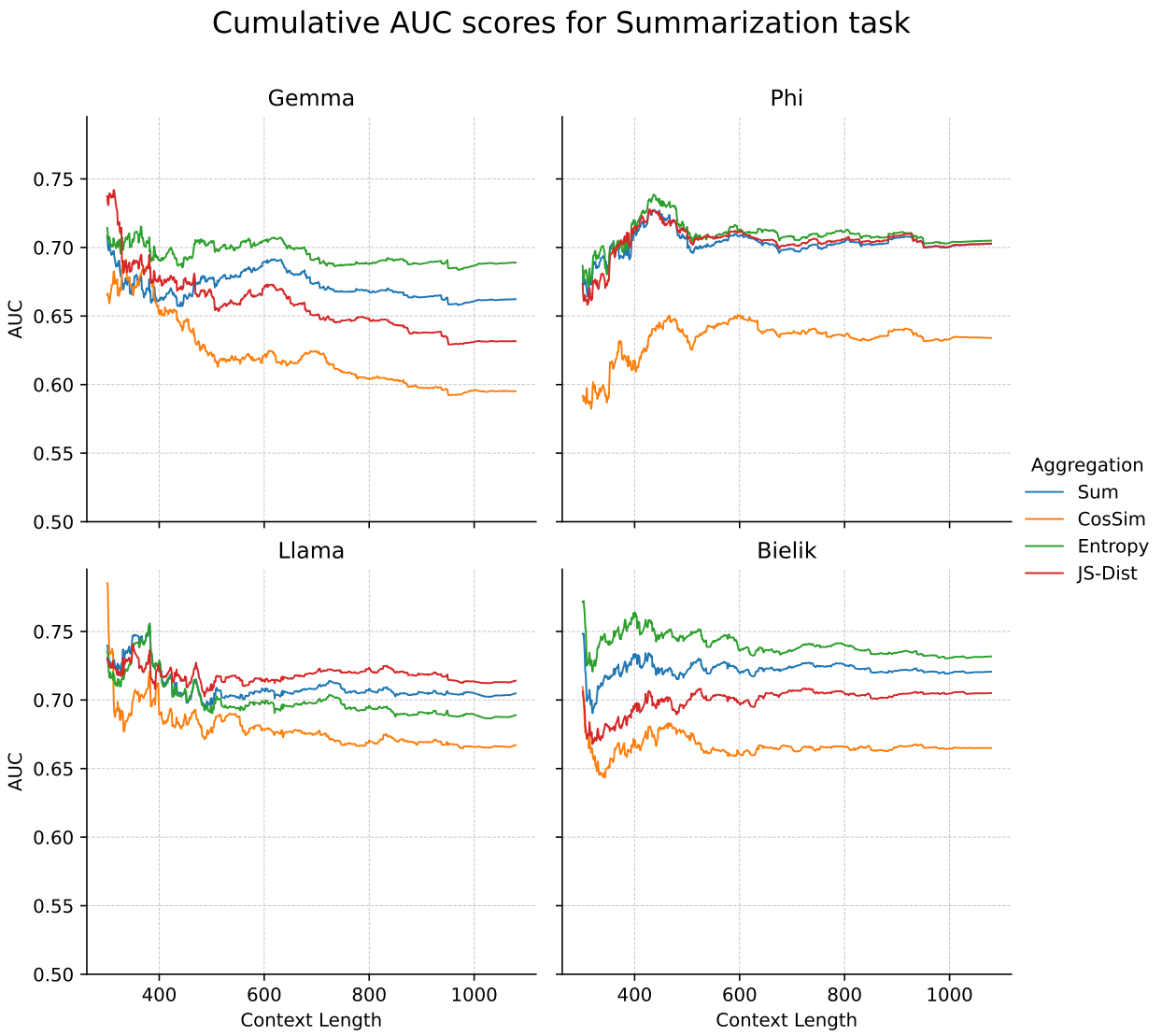

Figure 6.3: Cumulative AUC scores across context length for Summarization task (CNN/DM → XSum).

- **Llama**, however, exhibits the opposite trend: JS-Dist remains stable, but other features like Sum and Entropy degrade as context grows, suggesting difficulty in maintaining attention focus over longer sequences.
- **Bielik** and **Phi** show only mild sensitivity, with no consistent preference for short or long contexts.

These findings reinforce that while attention aggregation features are generally robust to input length, certain methods and models demonstrate context-length-specific dynamics. In particular, model-specific attention behaviors likely interact with feature design in ways that can either amplify or dilute the signal.

Crucially, the relative robustness of Sum and Entropy across configurations positions them as strong candidates for practical deployment in RAG systems, where retrieved context length can vary considerably. The relative instability of some features under length variation also emphasizes the value of task- and model-aware evaluation when designing hallucination detectors.

6.3. Comparison with Lookback Lens

This section compares the performance of the proposed attention aggregation methods against the Lookback Lens baseline, focusing on the Llama-2-7B-chat model. Figure 6.4 summarizes the AUC scores across QA and summarization task transfer settings.

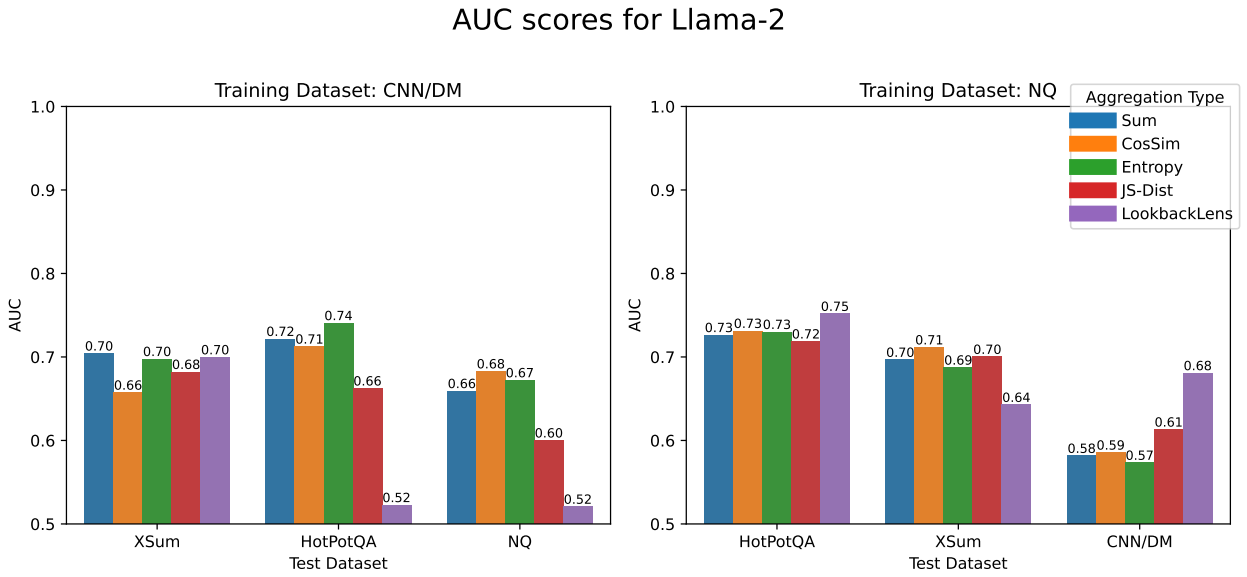

Figure 6.4: AUC comparison between aggregation methods and Lookback Lens on Llama-2 across QA and summarization transfer pairs. AUC = 0.5 is treated as a natural baseline (random classifier).

The Lookback Lens method demonstrates competent performance in in-domain scenarios. However, in cross-task configurations, its performance drops sharply—approaching random

guessing (AUC ≈ 0.52) in setups such as CNN/DM $\rightarrow$ NQ. This suggests a limited capacity for generalization beyond the task it was trained on.

In contrast, the proposed aggregation features show stronger cross-task robustness:

- When trained on summarization and evaluated on QA, **Sum** and **Entropy** significantly outperform Lookback Lens, indicating their ability to generalize across differing task structures.
- In QA-to-summarization transfer, **JS-Dist** and **CosSim** provide better performance.
- Lookback Lens retains moderate strength in specific setups when trained on NQ, highlighting that it can succeed when the generation patterns and alignment assumptions are preserved.

These results suggest that Lookback Lens is brittle in settings where input-output alignment is loose or task semantics change. In contrast, aggregation methods that synthesize global attention signals, particularly Sum and Entropy, retain their effectiveness across domains.

Overall, this comparison demonstrates that while Lookback Lens can perform competitively in narrow, task-aligned conditions, the proposed aggregation techniques offer superior versatility, making them more suitable for open-domain RAG applications where task diversity and distribution shift are common.

6.4. Cross-Task and Cross-Lingual Generalization

This section evaluates the generalization capability of the aggregation-based features under three transfer settings: same-task, cross-task, and cross-lingual. Each setup involves training on one dataset and testing on a different one, either within the same task family or across tasks or languages. The goal is to assess whether the proposed features can retain discriminative power under distributional shift.

6.4.1. Same-Task Transfer

Figure 6.5 presents the AUC scores for models trained and tested on different datasets within the same task category (QA or summarization).

Across most models, **Entropy** and **Sum** achieve the highest and most stable AUC scores. This suggests that these features generalize well across datasets that share similar task formats and content structures. Their ability to summarize overall attention strength (Sum) or quantify uncertainty (Entropy) appears robust to dataset-specific variations.

CosSim, by contrast, consistently underperforms, indicating a limited ability to model meaningful semantic relationships in attention distributions. **JS-Dist** performs moderately but is less stable across LLMs, implying sensitivity to specific attention behaviors or dataset semantics.

These patterns suggest that some aggregation features capture domain-invariant grounding signals, while others are more brittle to subtle dataset shifts even within a single task category.

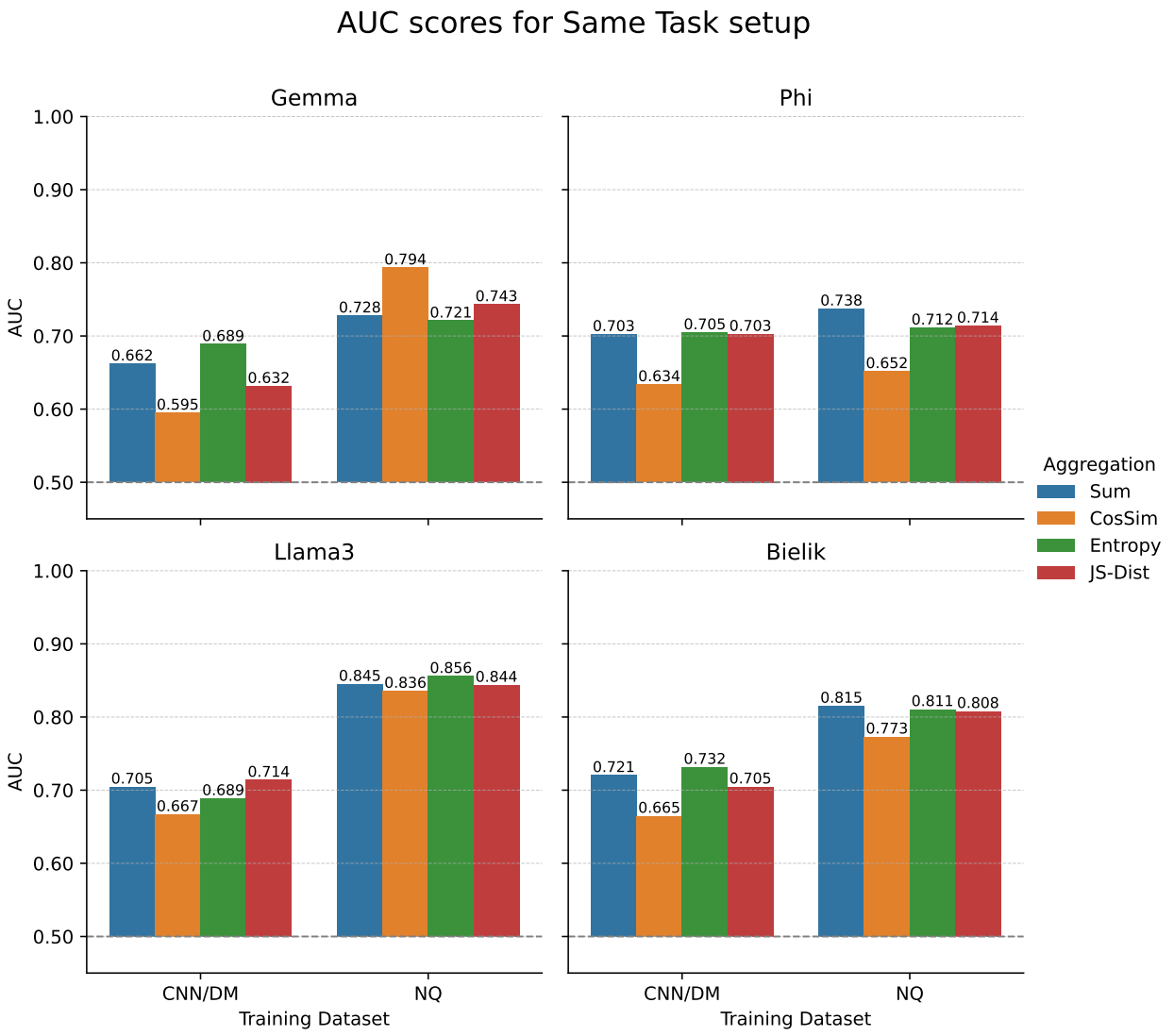

Figure 6.5: Same-task transfer AUC scores across LLMs.

6.4.2. Cross-Task Transfer

Figure 6.6 reports the performance when training on one task type (e.g., QA) and testing on another (e.g., summarization).

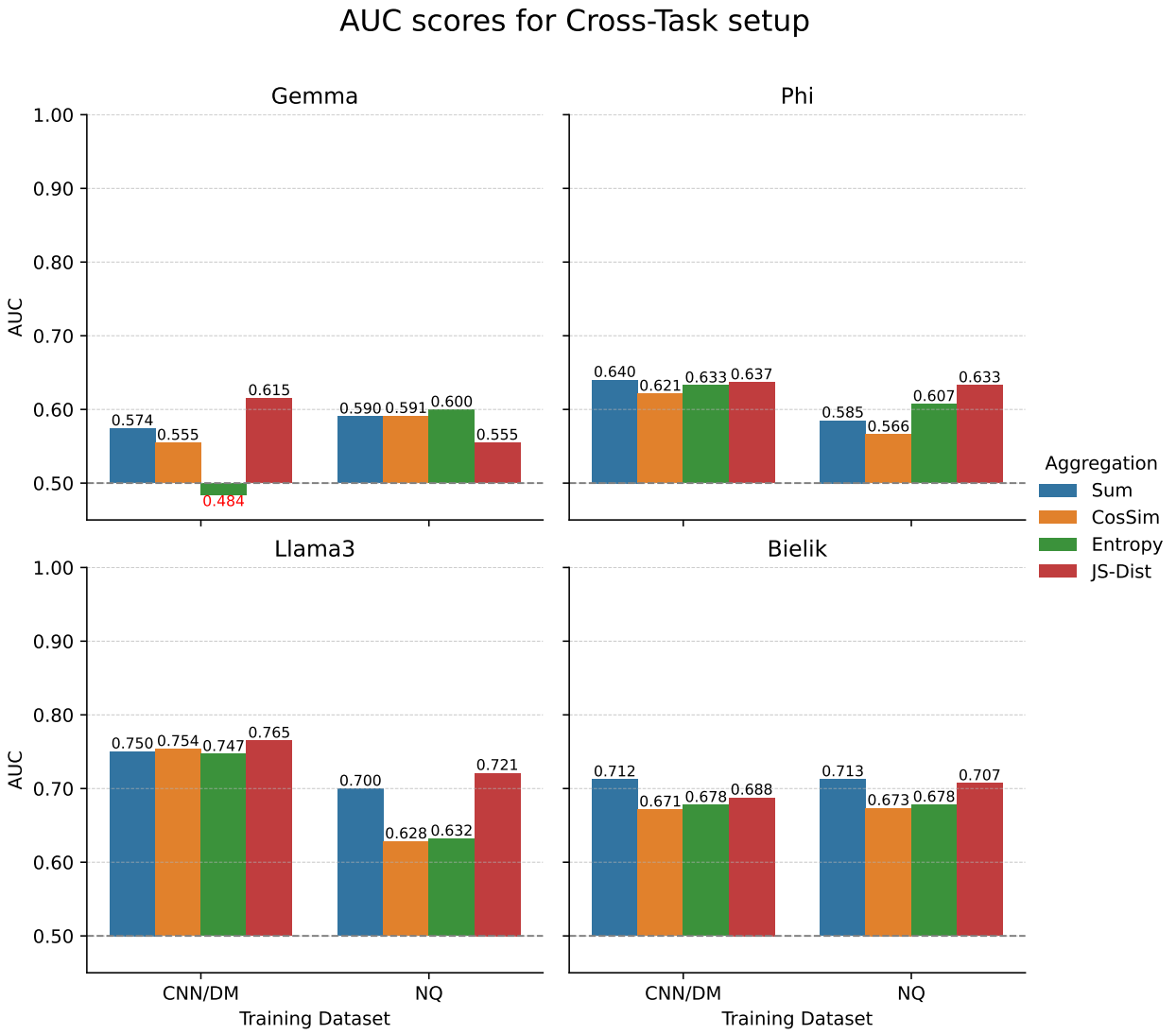

Figure 6.6: Cross-task transfer AUC scores.

Cross-task transfer poses a greater challenge due to differences in how information is structured, presented, and attended to. Among the features, **JS-Dist** exhibits the strongest generalization in most configurations, particularly when moving from summarization to QA. Its success may reflect its sensitivity to shifts in the divergence of attention across tokens, which can serve as a proxy for the model's uncertainty under unfamiliar generation formats.

Sum performs competitively, especially in setups where the grounding structure between tasks remains moderately aligned. However, **Entropy** becomes less consistent in this scenario, indicating that uncertainty-based signals may be less reliable when task semantics diverge.

Once again, **CosSim** yields the weakest results, reinforcing its limited capacity to encode

transferable grounding information across tasks.

6.4.3. Cross-Lingual Transfer

Figure 6.7 shows results for models trained on English datasets and tested on Polish counterparts, representing a realistic cross-lingual evaluation.

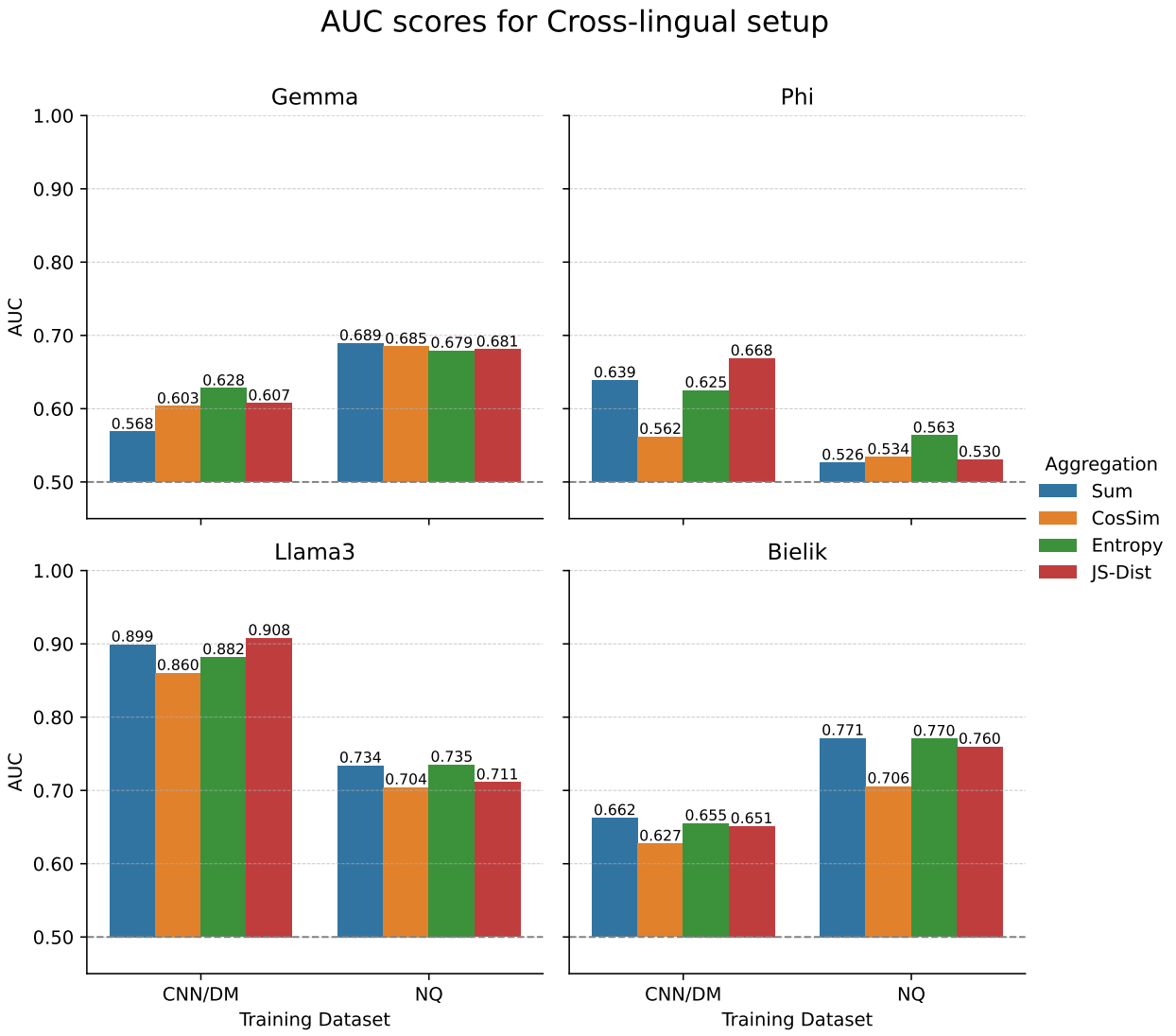

Figure 6.7: Cross-lingual generalization: AUC scores on Polish test sets.

In cross-lingual settings, **Entropy** and **JS-Dist** again perform best, with Entropy slightly more stable across language models. This suggests that both attention uncertainty and inter-token divergence capture signals that are to some extent language-agnostic. These features appear effective at tracking attention grounding regardless of surface form variation or language-specific tokenization.

Sum maintains moderate performance, though it is less consistent, possibly due to language-specific variations in attention spread. **CosSim**, as in other experiments, fails to

generalize and performs poorly.

6.4.4. Aggregated Ranking Summary

To synthesize results across all transfer setups, Table 6.1 reports the average rank and mean normalized AUC for each aggregation method.

Table 6.1: Aggregated performance across evaluation setups.

Setup	Aggregation	Avg. Rank	Mean Norm. AUC
Same-Task	Entropy	2.0	0.760
	Sum	2.0	0.753
	JS-Dist	2.6	0.575
	CosSim	3.4	0.200
Cross-Task	JS-Dist	2.0	0.672
	Sum	2.2	0.699
	CosSim	2.9	0.373
	Entropy	2.9	0.339
Cross-Lingual	Entropy	2.1	0.691
	Sum	2.4	0.599
	JS-Dist	2.4	0.566
	CosSim	3.1	0.321

The aggregated ranking confirms that **Sum** and **Entropy** are the most consistently high-performing features across same-task and cross-lingual settings, while **JS-Dist** is most effective in cross-task transfer. These distinctions suggest that different aggregation strategies excel under different types of generalization pressure.

Importantly, none of the features dominate across all setups, highlighting the complementary nature of their underlying signals. These findings support a hybrid approach or task-specific selection strategy when applying attention-based hallucination detection in practice.

6.5. Supervision with Semantic Entropy

6.5.1. Entropy Distributions

Figure 6.8 shows that semantic entropy values differ only marginally between hallucinated and non-hallucinated examples. Overlapping distributions indicate limited discriminative power.

Table 6.2 reports Cohen's κ agreement between hallucination labels and entropy-based labels. Only NQ achieves fair agreement; CNN/DM results are close to random. This confirms the weakness of semantic entropy as a standalone hallucination proxy.

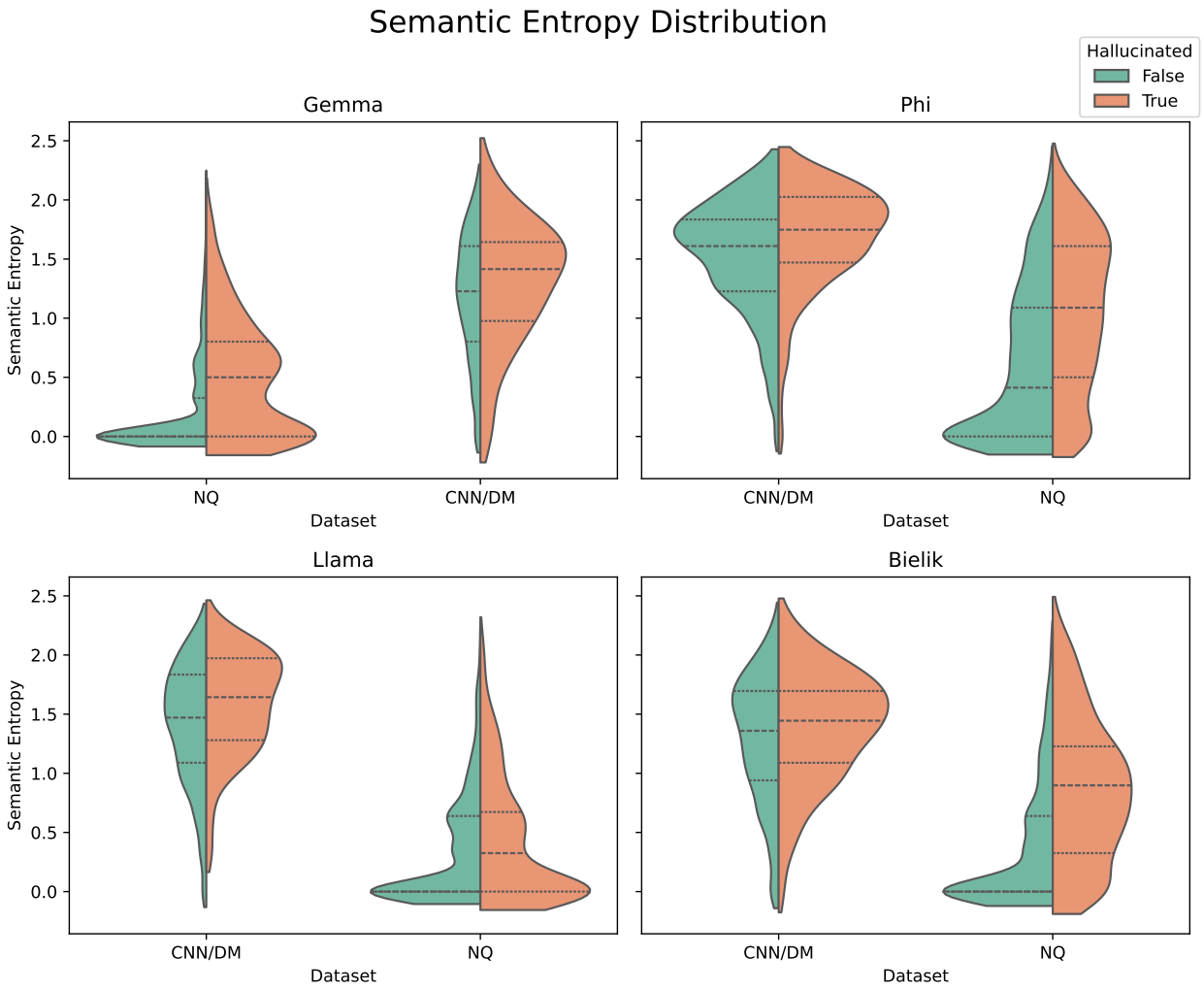

Figure 6.8: Distribution of semantic entropy by hallucination label.

Table 6.2: Agreement between hallucination labels and semantic entropy-based labels.

Dataset	**Model**	**Cohen's κ**
CNN/DM	Bielik	0.037
	Gemma	0.024
	Llama	0.044
	Phi	0.098
NQ	Bielik	0.306
	Gemma	0.233
	Llama	0.054
	Phi	0.266

6.5.2. Supervised Performance Using Entropy Targets

Before-Generation Token (BGT)

Figure 6.9 and Figure 6.10 show performance using semantic entropy as training supervision. The results decline significantly for most models. Summarization scores are near random.

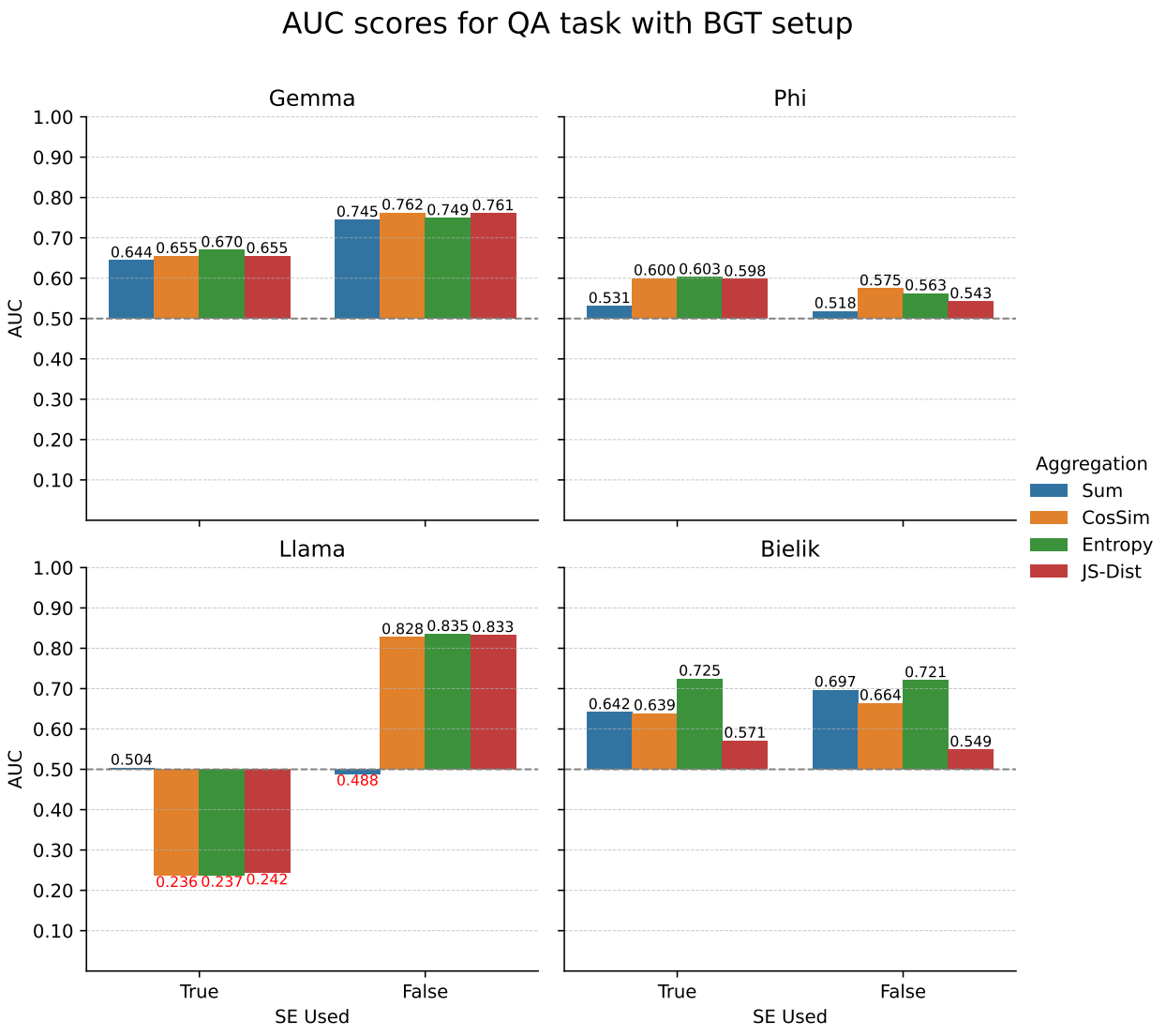

Figure 6.9: AUC scores using semantic entropy supervision for QA task and BGT setup.

Second-to-Last Token (SLT)

Figure 6.11 and Figure 6.12 confirm similar degradation under the SLT setup. Most models perform worse with entropy labels than with binary labels.

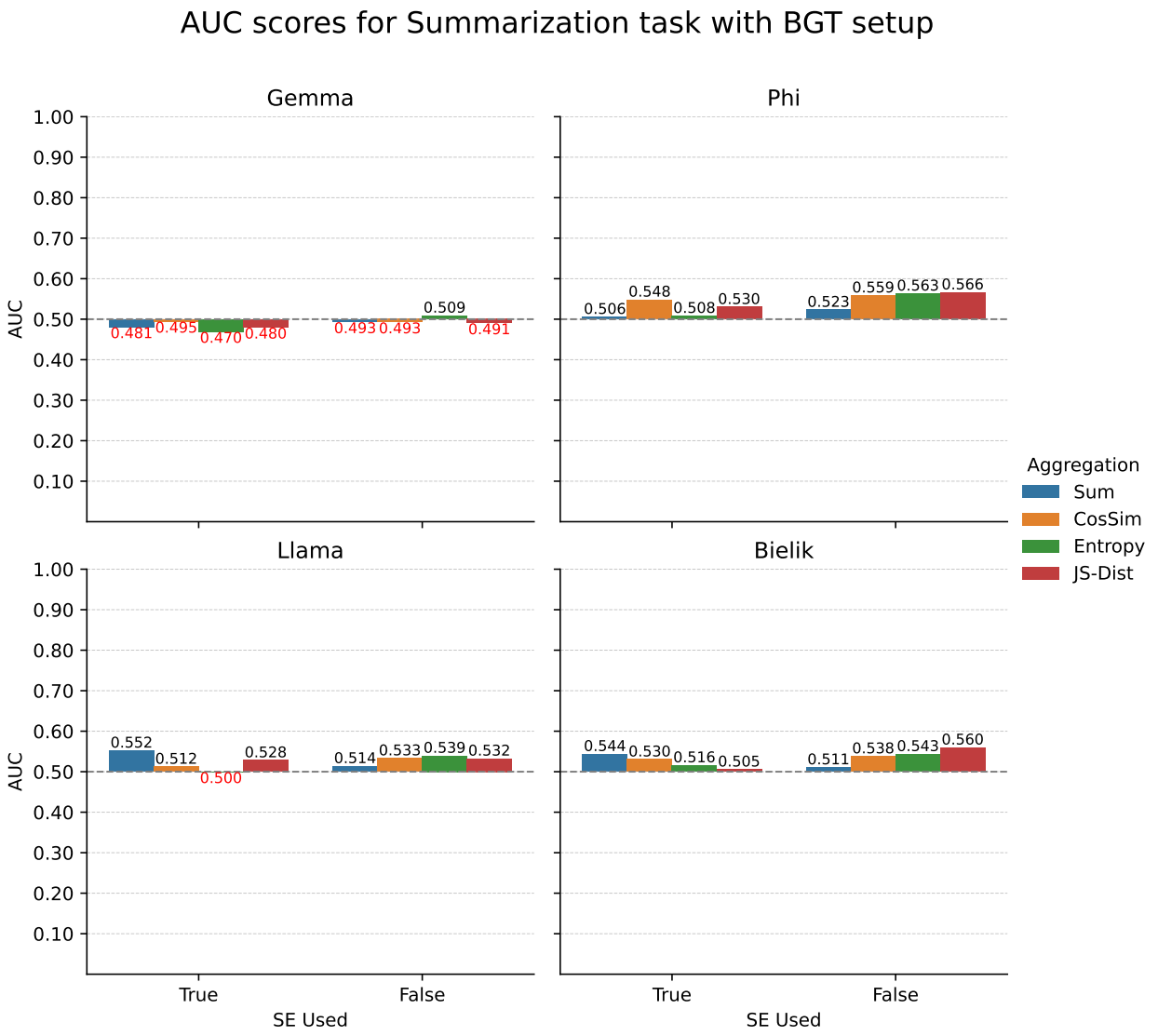

Figure 6.10: AUC scores using semantic entropy supervision for Summarization task and BGT setup.

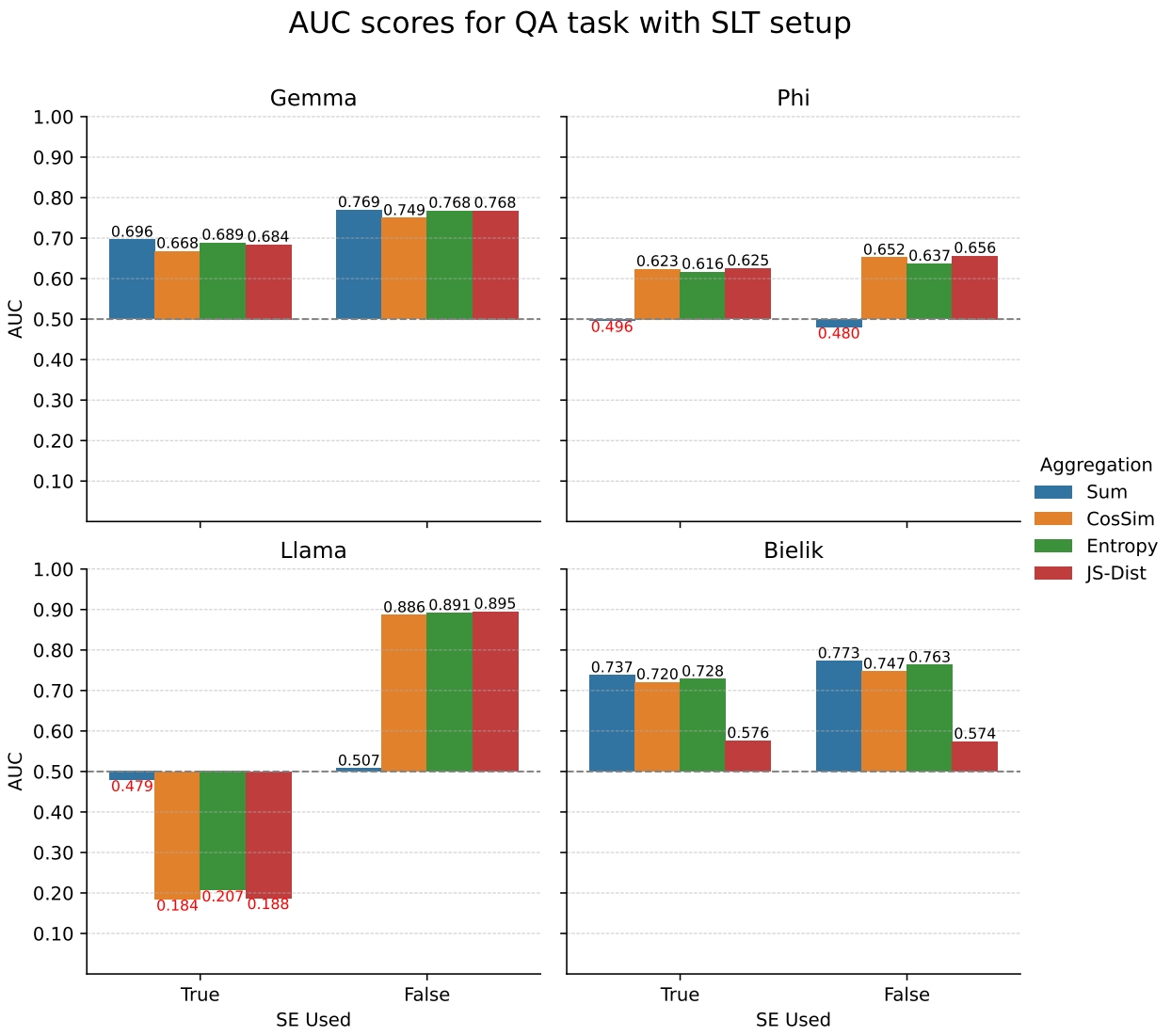

Figure 6.11: AUC scores using semantic entropy supervision for QA task and SLT setup.

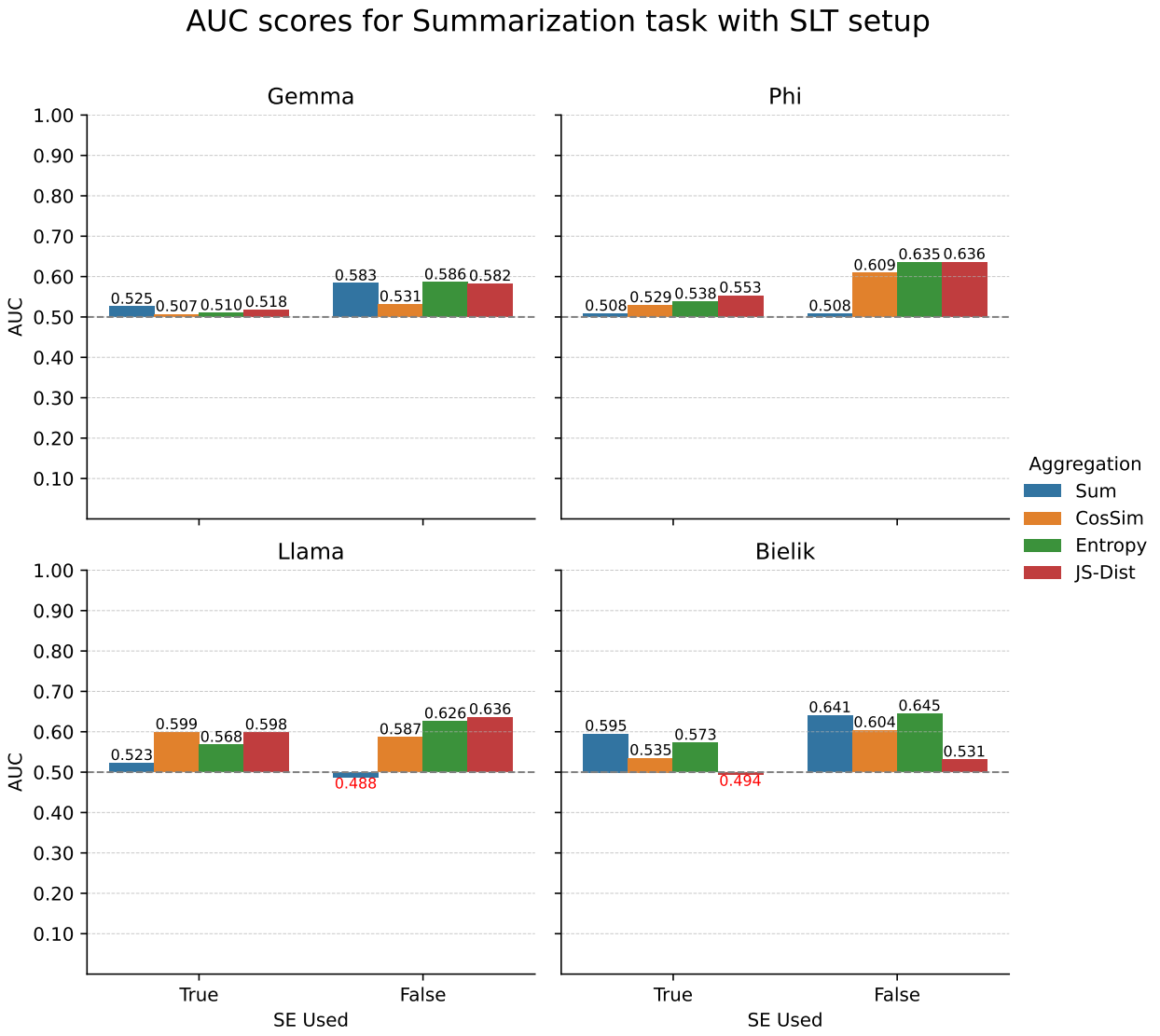

Figure 6.12: AUC scores using semantic entropy supervision for Summarization task and SLT setup.

6.6. Summary of Findings

The empirical results demonstrate that attention-based aggregation features offer a robust and interpretable alternative to existing hallucination detection techniques, particularly in cross-task and cross-lingual settings. Among the proposed methods, Sum and Entropy consistently yield the highest AUC scores in same-task and multilingual scenarios, suggesting their ability to capture stable grounding signals across diverse language models and datasets. These features appear resilient to variations in input context length and dataset structure, reinforcing their practical utility in real-world RAG pipelines.

JS-Dist, while more variable in performance, proves particularly effective under cross-task transfer conditions. This suggests that distributional divergence in attention patterns may capture task-invariant grounding signals, albeit with some sensitivity to model-specific attention dynamics. In contrast, CosSim underperforms across all configurations, indicating limited utility as a standalone signal for grounding or hallucination.

When compared with the Lookback Lens baseline, the aggregation methods, especially Sum, Entropy, and JS-Dist, demonstrate significantly better generalization across tasks. Lookback Lens performs well within QA tasks but fails in cross-domain or summarization scenarios, underscoring its brittleness under distributional shift.

Finally, experiments involving semantic entropy as a weak supervisory signal reveal its limitations. Although interpretable and model-internal, semantic entropy shows a weak correlation with hallucination labels and often leads to degraded classification performance when used as a substitute for binary supervision. This reinforces the value of direct, interpretable grounding features over indirect uncertainty measures.

Overall, the results confirm the viability of lightweight, interpretable attention aggregations as effective tools for hallucination detection. Each method exhibits distinct strengths under different generalization pressures, suggesting that a task-aware or ensemble approach may yield further performance gains.

7. Conclusions

This thesis addressed the challenge of detecting contextual hallucinations in RAG systems by using internal attention signals from LLMs. Specifically, it introduced a set of attention aggregation techniques designed to extract lightweight, interpretable features from attention maps, and evaluated their effectiveness against strong baselines including Lookback Lens and Semantic Entropy.

The proposed aggregation methods, Sum, Cosine Similarity, Entropy, and Jensen-Shannon Distance, were applied over decoder-only architectures to model the alignment between generated tokens and retrieved evidence. These features were then used to train simple supervised classifiers for token-level hallucination detection.

Key Findings

The evaluation spanned six datasets (four English, two Polish), five LLMs (Llama-2, Llama-3, Gemma, Phi, and Bielik), and a diverse set of experimental setups. The main findings can be summarized as follows:

- **Attention aggregation enables effective hallucination detection.** The proposed Sum and Entropy methods consistently performed well across datasets and models, often outperforming or matching the Lookback Lens baseline, especially when trained on summarization data.
- **Lookback Lens is brittle in transfer settings.** While effective within QA tasks, Lookback Lens showed poor generalization to other task types, underscoring its sensitivity to prompt structure and training distribution.
- **Jensen–Shannon Distance generalizes well but lacks stability.** JS-Dist offered the strongest cross-task and cross-lingual transfer performance, but exhibited greater variance, occasionally underperforming relative to simpler methods.
- **Cosine Similarity is a weak predictor.** Across nearly all conditions, CosSim ranked lowest, suggesting that agreement among heads does not strongly correlate with factual grounding.
- **Context length is not a major driver of hallucination.** Model performance remained stable across input context sizes, indicating that attention-based features are robust to variability in passage length.
- **Semantic entropy is not a reliable supervision signal.** Despite its theoretical appeal, semantic entropy failed to distinguish hallucinated from grounded outputs effectively. Training classifiers on entropy-derived labels consistently underperformed relative to ground-truth binary labels.

Contributions

This work makes several novel contributions to the study of factual consistency and hallucination detection in LLMs:

- **Methodological:** Introduction of four attention-based aggregation strategies for hallucination detection that are model-internal, interpretable, and computation-efficient.
- **Empirical:** A large-scale evaluation across five LLMs, six datasets, and multiple languages and task types, including a rigorous comparison with state-of-the-art baselines.
- **Analytical:** Insights into the robustness of attention patterns under context variation and the limits of semantic entropy as a supervision mechanism.
- **Practical:** An end-to-end detection pipeline that integrates seamlessly with LLM inference and can be extended to real-time hallucination detection in RAG systems.

Limitations and Future Work

Although this thesis demonstrates the promise of attention-based aggregation for hallucination detection, several limitations remain:

- **Task scope:** Only QA and summarization were studied. Future work could extend the approach to dialogue, reasoning, or code generation, where hallucination manifests differently.
- **Classifier simplicity:** Logistic regression was used for its interpretability and efficiency. More expressive classifiers, such as MLPs or lightweight probes, can capture richer attention patterns without sacrificing interpretability.
- **Integration with decoding-time mitigation:** The current framework is used purely for post hoc detection. Future work could explore its integration into the decoding loop to dynamically steer or constrain generation in real time, potentially reducing hallucination as it emerges.

Final Remarks

This thesis demonstrates that hallucination detection in RAG systems can be addressed through the principled use of internal model signals, without requiring costly external retrievers or opaque black-box metrics. The proposed attention aggregation techniques are interpretable, lightweight, and extensible, offering a promising direction for the scalable and reliable deployment of language models.

As LLMs continue to power high-stakes applications across domains and languages, tools for detecting and diagnosing ungrounded generation will become increasingly vital. This work contributes a modular foundation for such tools and lays the foundation for future research in model transparency, controllability, and factual alignment.

Bibliography

[1] S. Abnar and W. Zuidema. Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] M. AI. Introducing llama 3: Meta's most capable open models yet. https://ai.meta.com/blog/meta-llama-3/, 2024. Accessed: 2025-05-26.

[3] A. L. at Wrocław University of Science and Technology. Bielik: A large language model for polish. https://huggingface.co/ai-lab-PL/bielik-11b-v2.3-instruct, 2024. Accessed: 2025-05-26.

[4] A. Azaria and T. Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2310.01391*, 2023.

[5] J. Binkowski, D. Janiak, A. Sawczyn, B. Gabrys, and T. Kajdanowicz. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. To appear.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

[7] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[8] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 276–286, 2019.

[9] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022.

[10] G. DeepMind. Gemma 2: Open models by google deepmind. https://ai.google.dev/gemma, 2024. Accessed: 2025-05-26.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[12] L. Gao, A. Bakhtin, N. Scales, L. Zettlemoyer, N. Houlsby, and M. Lewis. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[13] Q. Han, D. Khashabi, X. V. Lin, and D. Roth. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*, 2024.

[14] P. He, X. Liu, J. Gao, and W. Chen. Debertav2: Improving deberta using larger training data and training techniques. *arXiv preprint arXiv:2111.09543*, 2021.

[15] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend, 2015.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[17] L. Huang and et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[18] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556, 2019.

[19] Z. Ji, N. Lee, J. A. Fries, T. H. Yu, I. Hsu, S. Whitmore, T. Xia, H. Zhang, D. Sengupta, S. Amershi, et al. A survey of hallucination in natural language generation. *arXiv preprint arXiv:2301.12022*, 2023.

[20] J. Kossen, J. Han, M. Razzak, L. Schut, S. Malik, and Y. Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024.

[21] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, and S. Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

[23] N. F. Liu, D. Yu, T. Zhang, Y. Chien, A. Bosselut, and N. A. Smith. Gpteval: Benchmarking large language models as graders for natural language generation. *arXiv preprint arXiv:2306.05685*, 2023.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[25] N. Manakul and M. J. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Transactions of the Association for Computational Linguistics*, 11:621–638, 2023.

[26] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989.

[27] P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.

[28] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[29] M. Ogrodniczuk and M. Łaziński. The polish corpus of summaries. In *Proceedings of the 5th Language and Technology Conference (LTC)*, Poznań, Poland, 2011.

[30] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, D. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2463–2473, 2019.

[31] M. Research. Phi-3 technical report. https://www.microsoft.com/en-us/research/project/phi-3/, 2024. Accessed: 2025-05-26.

[32] K. Shuster, M. Komeili, and J. Weston. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2104.07567*, 2021.

[33] TODO. Aggtruth: Contextual hallucination detection using aggregated attention scores in llms. In *International conference on computational science*, page (accepted for publication). Springer, 2025.

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, N. Goyal, P. Batra, P. Mazaré, H. Jégou, and A. Joulin. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[35] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[37] E. Voita, J. Talbot, R. S. Moiseev, and I. Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.

[38] J. Wang, H. Zhang, Y. Liu, D. Zhang, and Y. Wu. Detecting hallucinations in large language models using semantic entropy. *Nature*, 619:112–120, 2024.

[39] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[40] F. Wawrzyniak, M. Plewczyński, M. Piasecki, and M. Perełkiewicz. Polqa: Polish reading comprehension dataset for question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2274–2286, Dubrovnik, Croatia, 2023.

[41] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, 2019.

[42] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018.

[43] R. Zhong, P. Lewis, A. Dai, and A. W. Yu. How does in-context learning help hallucination in open-domain qa? *arXiv preprint arXiv:2210.07128*, 2022.

List of Figures

List of Tables