Master's Thesis

Preemptive Hallucination Control in LLMs Using Representation Engineering with Uncertainty Aware Steering

OBULE Peter Kasobi

Supervisor NAOYA Inoue

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2026

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Associate Professor Naoya Inoue, for his guidance, insightful feedback, and continued support throughout this research.

I am also grateful to my lab mates for their discussions, encouragement, and collaborative environment, as well as to my family for their constant support and understanding.

Finally, I acknowledge the Japan Advanced Institute of Science and Technology (JAIST) for providing the academic environment that made this work possible.

Abstract

Large Language Models (LLMs) exhibit a persistent tendency to generate fluent yet factually incorrect outputs, commonly referred to as hallucinations. This failure mode remains prevalent despite advances in model scaling, instruction tuning, and alignment, posing significant risks in high-stakes applications. Recent work has shown that hallucinations can be mitigated at inference time by intervening directly in a model's internal representations, without retraining or external knowledge sources. In particular, representation engineering (RepE) demonstrates that behavioral properties such as truthfulness correspond to approximately linear directions in activation space, and that injecting these directions during inference can bias model outputs toward factual correctness.

However, existing inference-time steering methods apply interventions with a fixed strength across all tokens and decoding steps, implicitly assuming that hallucination risk is uniform throughout generation or decision-making. This assumption ignores substantial variation in model uncertainty across tokens and answer options. While token-level uncertainty estimation has been extensively studied as a diagnostic tool for calibration and hallucination detection, its use as a continuous control signal for inference-time intervention remains poorly understood.

This thesis investigates whether model-internal uncertainty can be used to adaptively modulate inference-time representation steering. We introduce *Uncertainty-Aware Scaling* (UAS), a framework that dynamically adjusts the magnitude of a fixed representation-level steering direction based on token-level uncertainty signals computed during inference. UAS is evaluated using multiple uncertainty estimators, including predictive entropy, Mahalanobis distance-based uncertainty, and recurrent attention-based uncertainty, and is applied selectively to the final transformer layers where interventions most strongly influence output decisions.

Experiments are conducted on TruthfulQA and CREAK, using both multiple-choice and generative evaluation protocols across several open-source LLMs. Results show that static representation steering consistently improves factual decision-making across models, confirming its robustness as an inference-time hallucination mitigation technique. In contrast, uncertainty-aware scaling exhibits mixed and strongly model-dependent behavior. For some configurations, uncertainty modulation provides modest gains, while in many cases it attenuates effective steering or degrades performance.

Detailed analysis reveals that these outcomes are driven by weak alignment between token-level uncertainty signals and incorrect content at the layers where steering is applied. As a result, uncertainty-aware modulation often operates in a noise-dominated regime unless the base steering signal is sufficiently strong. In free-form generation, uncertainty-aware scaling consistently degrades truthfulness and stability, indicating that token-level uncertainty is ill-suited as a control signal for generative decoding.

Overall, this work clarifies both the potential and the limitations of uncertainty-driven inference-time control. While uncertainty-aware modulation can refine representation steering under specific conditions, static representation engineering remains the most robust and broadly effective inference-time intervention for hallucination mitigation in large language models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Large Language Models (LLMs), including Generative Pre-trained Transformer (GPT) [26] and the Large Language Model AI (LLaMA) models [29], have demonstrated remarkable performance across a wide range of natural language processing tasks such as question answering and text generation. Despite these advances, a persistent limitation [12] of such models is their tendency to produce outputs that are fluent and confident yet factually incorrect or unsupported by evidence. This phenomenon poses serious risks in high-stake domains including healthcare, law, education, and scientific research, where erroneous information can lead to harmful or misleading outcomes.

In the context of large language models, *hallucinations* refer to the generation of fluent and confident statements that are factually incorrect, unverifiable, or unsupported by reliable evidence. Unlike simple syntactic or grammatical errors, hallucinations arise when a model assigns high probability to plausible-looking continuations that are not factual. This behavior is a direct consequence of the probabilistic sequence modeling paradigm underlying modern LLMs, in which models are trained to maximize the likelihood of next-token predictions conditioned on context, rather than to explicitly reason about factual correctness or evidential grounding [19, 13]. Because likelihood-based training rewards statistical plausibility over grounded truth, models may confidently reproduce misconceptions, fabricate details, or overgeneralize from spurious correlations present in the training data. This effect is particularly pronounced in settings involving rare or underrepresented facts, ambiguous prompts, or under-specified contexts, where the training distribution provides weak or conflicting signals [13, 15]. Moreover, instruction fine-tuning can further exacerbate hallucinations by encouraging models

to provide answers even when uncertainty or abstention would be more appropriate [12].

A variety of approaches have been proposed to mitigate hallucinations in large language models. Supervised fine-tuning and reinforcement learning from human feedback (RLHF) aim to align model outputs with human judgments of correctness, but require substantial annotated data and may not generalize beyond the domains or distributions on which they are trained [25]. Retrieval-augmented generation (RAG) incorporates external knowledge sources to ground model outputs, improving factuality in many settings but introducing additional system complexity, latency, and dependency on retrieval quality [17]. Other methods [32, 11, 6, 14] encourage abstention or uncertainty expression through prompt engineering or post-hoc calibration, yet these techniques often rely on heuristic thresholds and do not directly alter the model's internal decision process [15].

More recently, inference-time intervention methods have emerged as a lightweight alternative that seeks to influence model behavior without updating parameters or requiring external tools. Representation-level approaches, in particular, intervene directly in hidden activations to bias the model toward desired behaviors such as honesty or factual correctness [33, 18]. While effective, most existing inference-time steering methods apply static interventions uniformly across tokens and decoding steps, implicitly assuming that the model's propensity to hallucinate is constant throughout generation or decision-making. This assumption overlooks the substantial variation in model uncertainty across tokens, motivating approaches that adapt intervention strength dynamically based on the model's uncertainty.

Despite these advances, benchmark evaluations reveal that hallucinations remain a persistent failure mode even under modern alignment and inference-time control techniques. Datasets such as TruthfulQA [19] and CREAK [24] provide concrete evidence that hallucinations persist even in large, instruction-tuned language models. TruthfulQA shows that models trained on internet-scale corpora systematically reproduce common misconceptions when asked direct questions, often assigning higher confidence to plausible but incorrect answers than to factual ones. CREAK complements this finding by framing hallucination as a factual verification problem, requiring models to distinguish between true and false declarative claims under a controlled multiple-choice setting. Together, these benchmarks demonstrate that hallucinations are not limited to free-form generation, but also arise in structured decision-making tasks where models must select among competing factual alternatives. This makes them well suited for analyzing inference-time interventions that aim to reshape internal confidence and selection behavior without retraining.

1.2 Problem Statement

Despite extensive progress in hallucination mitigation, existing approaches typically rely on additional training, external knowledge sources, or heuristic post-processing. While effective in specific settings, such methods increase system complexity, introduce new failure modes, or require curated data that may not generalize across tasks and domains.

Inference-time control has emerged as a promising alternative, offering a lightweight mechanism for influencing model behavior without modifying parameters or introducing external dependencies. Representation-level steering methods, in particular, demonstrate that targeted interventions in hidden activations can bias model outputs toward desired behaviors [33, 18] such as truthfulness or factual consistency. However, current inference-time steering techniques are predominantly static: they apply a fixed intervention uniformly across all tokens and decoding steps.

This static assumption overlooks a fundamental property of large language models: internal confidence and uncertainty vary substantially across tokens, answer options, and decision points. In structured decision-making tasks such as multiple-choice question answering, some tokens or options are associated with high epistemic uncertainty, while others are generated with strong confidence. Uniform steering therefore risks being poorly calibrated—either insufficient to influence uncertain decisions or overly aggressive in confident regions, potentially degrading accuracy or informativeness.

The central problem addressed in this thesis is that existing inference-time representation steering methods lack a principled mechanism for adapting intervention strength to the model's internal uncertainty. While token-level uncertainty signals such as predictive entropy, confidence scores, and variance estimates have been extensively studied as diagnostic tools for calibration, hallucination detection, and selective prediction [15, 13, 6], their use as continuous, token-level modulators of representation-level inference-time control remains limited and poorly understood. In particular, it is unclear whether dynamically scaling representation-level interventions based on uncertainty can reliably improve factual decision-making, or under what conditions such scaling may instead introduce noise or weaken effective control.

This thesis investigates this gap by systematically evaluating uncertainty-aware inference-time representation steering across multiple uncertainty estimation strategies, model architectures, and factual reasoning benchmarks.

1.3 Research Questions

To address the limitations of static inference-time representation steering identified in the problem statement, this thesis investigates whether model-internal uncertainty can be used to adaptively modulate inference-time interventions in a principled and effective manner.

The study is guided by the following research questions:

1. **Primary Research Question:** Under what conditions does uncertainty-aware, token-level inference-time representation steering improve factual decision-making in large language models compared to static steering?

2. **Secondary Research Question:** What trade-offs arise between truthfulness and informativeness when uncertainty-aware inference-time steering is applied?

1.4 Research Aim

The aim of this research is to evaluate whether uncertainty-aware, token-level inference-time representation steering can improve factual decision-making in large language models compared to static steering, and to identify the conditions under which such uncertainty-aware modulation is beneficial or harmful.

1.5 Overview of Approach

This thesis investigates hallucination mitigation through inference-time interventions that directly modify a model's internal representations during prediction, without updating model parameters or relying on external knowledge sources.

At a high level, the approach consists of two stages. First, a representation-level intervention is constructed that biases the model toward truthful behavior. Second, this intervention is either applied uniformly (static steering) or modulated dynamically based on the model's internal uncertainty (uncertainty-aware steering).

Inference-time representation steering. The core intervention mechanism builds on prior work in representation engineering [33], which demonstrates that certain behavioral attributes—such as truthfulness or honesty—can

be associated with directions in a model's hidden activation space. These directions are estimated by contrasting internal representations elicited by truthful and untruthful prompts or labeled answer pairs. The resulting *steering vector* is then injected into selected transformer layers during inference, nudging the model's internal states toward truth-consistent representations.

In its simplest form, this intervention is applied with a fixed strength across all tokens and decoding steps. This static variant serves as a strong baseline and corresponds to standard inference-time representation steering (RepE).

Uncertainty-aware scaling. Building on this baseline, the proposed method introduces *Uncertainty-Aware Scaling* (UAS), which allows the strength of the representation-level intervention to vary across tokens. Instead of applying a uniform steering magnitude, UAS modulates the intervention based on signals derived from the model's internal uncertainty at each decoding step.

Intuitively, this design aims to concentrate steering on decision points where the model is uncertain—such as when selecting between competing factual alternatives—while reducing intervention in confident regions where steering may be unnecessary or harmful. Several uncertainty signals are explored, including entropy-based uncertainty, distance-based measures in representation space, and attention-based uncertainty indicators. These signals are computed directly from the model's internal states during inference.

Evaluation strategy. The proposed methods are evaluated primarily in structured factual decision-making settings, where hallucinations manifest as incorrect answer selection rather than open-ended fabrication. Experiments are conducted on multiple-choice benchmarks such as TruthfulQA (MC1) and CREAK, which require models to select correct answers among plausible alternatives.

Performance is assessed using accuracy-based metrics as well as confidence-sensitive analyses that examine how interventions affect the model's relative preference between correct and incorrect options. Additional diagnostic analyses investigate how uncertainty-aware scaling interacts with intervention strength, layer selection, and model architecture.

Through this evaluation, the thesis aims to characterize not only whether uncertainty-aware steering improves factual decision-making, but also when and why such modulation succeeds or fails.

Chapter 2

Literature Review

2.1 Hallucinations in Large Language Models

Hallucinations in large language models (LLMs) refer to the generation of outputs that are fluent, coherent, and contextually appropriate, yet factually incorrect, unverifiable, or unsupported by reliable evidence. As LLMs are increasingly deployed in high-stakes domains such as healthcare, education, law, and scientific research, hallucinations have emerged as a central challenge to their reliability and safe deployment.

Prior work characterizes hallucination not as an isolated failure mode, but as a structural consequence of modern neural language modeling. Ji et al. [13] and Huang et al. [12] argue that hallucinations arise from the probabilistic training objective underlying LLMs: models are optimized to maximize the likelihood of observed text rather than to ensure factual correctness or epistemic validity. As a result, LLMs may assign high probability to outputs that are statistically plausible but factually false, especially when prompts are ambiguous, under-specified, or involve rare or disputed facts.

Empirical studies demonstrate that hallucinations persist even as model scale and training data increase. Lin et al. [19] show that larger language models trained on internet-scale corpora often reproduce common misconceptions with high confidence, indicating that scale alone does not resolve factual unreliability. At the same time, Kadavath et al. [15] find that language models can meaningfully estimate their own likelihood of being correct when explicitly prompted to report confidence, enabling selective prediction. However, this ability does not prevent models from generating fluent but incorrect answers under standard prompting, particularly in settings where plausible misconceptions compete with factual knowledge. Together, these findings suggest that while uncertainty-related signals can be informative

about model behavior, they are not automatically leveraged by language models during standard inference.

Hallucinations manifest across a range of task settings. In generative question answering and open-ended text generation, models may fabricate details, invent sources, or overgeneralize from spurious correlations present in the training data [22, 20]. In structured factual reasoning tasks, hallucinations instead appear as incorrect decisions under constrained answer formats. For example, TruthfulQA demonstrates that models often select plausible but incorrect answers over factual ones in multiple-choice settings [19], while CREAK frames hallucination as a binary factual verification problem, requiring models to judge whether a declarative claim is true or false [24]. Together, these benchmarks show that hallucinations are not limited to free-form generation, but also arise in structured factual decision-making contexts.

2.1.1 Types of Hallucination

Prior work categorizes hallucinations in large language models according to their relationship with the input context and external knowledge sources. A widely adopted taxonomy, summarized by Ji et al. [13], distinguishes between *intrinsic* and *extrinsic* hallucinations.

Intrinsic hallucinations occur when a model's output contradicts, distorts, or misrepresents information that is explicitly present in the input or source context [3, 1, 13, 27, 9]. This form of hallucination is commonly studied in tasks such as summarization and document-grounded question answering, where faithfulness to the provided input is critical. Empirical analyses show that even when relevant evidence is available, models may selectively omit, alter, or misinterpret salient facts [22, 20].

Extrinsic hallucinations, by contrast, arise when models introduce information that is not supported by either the input or external reality [1, 5]. These hallucinations are particularly prevalent in open-domain question answering and free-form generation tasks, where models must rely on parametric knowledge encoded during training rather than explicit grounding [13, 20]. In such settings, models may fabricate entities, causal explanations, or factual details that appear plausible but lack evidential support.

This distinction has proven useful for analyzing hallucination behavior across task settings and evaluation benchmarks. However, the boundary between intrinsic and extrinsic hallucinations is not always sharp: hybrid cases can arise when models both misinterpret available evidence and supplement it with unsupported information. Moreover, while this taxonomy provides a descriptive framework for categorizing errors, it does not by itself explain

when or why hallucinations occur during inference, nor how model-internal signals such as uncertainty evolve across different hallucination types.

2.2 Uncertainty Estimation in Large Language Models

Uncertainty estimation seeks to quantify a model's confidence in its predictions and has long been studied as a mechanism for identifying unreliable or error-prone outputs. In the context of large language models (LLMs), uncertainty has been proposed as a useful signal for detecting hallucinations, calibrating model confidence, and enabling selective prediction [13, 15].

Unlike Bayesian models, modern neural language models do not explicitly represent epistemic uncertainty. Instead, uncertainty must be inferred indirectly from output distributions or internal representations. As a result, uncertainty estimation in LLMs typically captures *predictive uncertainty*—reflecting ambiguity among plausible continuations—rather than true uncertainty about underlying facts or world knowledge [21, 10]. Despite this limitation, predictive uncertainty has been shown to correlate with model error and unreliability across a range of tasks.

2.2.1 Sequence-level Uncertainty Estimation

Sequence-level uncertainty estimation assigns a single confidence or uncertainty score to an entire generated output. Common metrics include average negative log-likelihood, perplexity, or normalized sequence probability. Such measures are widely used for response ranking, selective prediction, and post-hoc filtering of unreliable generations [15, 20].

While sequence-level uncertainty is computationally efficient and easy to interpret, it conflates multiple sources of variation, including sequence length, stylistic choices, and lexical diversity. More importantly, it provides no information about where uncertainty arises within a generated response. Empirical analyses of hallucinations suggest that factual errors often originate from localized decision points rather than uniformly across the entire sequence [13]. Consequently, global uncertainty measures may obscure the token-level dynamics that give rise to hallucinated content.

Mahalanobis distance-based uncertainty. Mahalanobis distance-based uncertainty quantifies how far a model's internal representation deviates from a reference distribution derived from in-domain data. Let $h \in R^d$ denote a

hidden representation and let (μ, Σ) denote the empirical mean and covariance estimated from reference representations. The Mahalanobis distance is defined as

$$D_M(h) = \sqrt{(h-\mu)^\top \Sigma^{-1}(h-\mu)}.$$

Originally introduced for out-of-distribution detection in neural networks [16], Mahalanobis distance has primarily been applied at the example or sequence level, using pooled or final-layer representations.

In the context of language models, Mahalanobis distance is typically computed over aggregated sequence representations. Its use as a token-level uncertainty signal is less explored and requires treating individual token representations as inputs to the distance computation. This adaptation is explored in this thesis and described in detail in Chapter 3.

2.2.2 Token-level Uncertainty Estimation

Token-level uncertainty estimation assigns uncertainty scores to individual generated tokens, enabling fine-grained analysis of confidence dynamics throughout the decoding process. This granularity is particularly relevant for hallucination analysis, as incorrect factual claims may emerge from a small subset of high-uncertainty tokens embedded within otherwise fluent outputs.

Predictive entropy. Given the model's output distribution over the vocabulary at token position t, denoted by $p(y_t \mid x, y_{<t})$, predictive entropy is defined as

$$H(y_t) = -\sum_{v \in \mathcal{V}} p(y_t = v \mid x, y_{<t}) \log p(y_t = v \mid x, y_{<t}),$$

where $\mathcal{V}$ is the vocabulary. Higher entropy indicates greater ambiguity among plausible next-token candidates and is commonly used as a token-level uncertainty measure in language models [21, 20].

Recurrent Attention-Based Uncertainty Quantification

Recurrent attention-based uncertainty quantification (RAUQ) estimates model uncertainty by analyzing the structure and temporal evolution of attention distributions within transformer models. The core intuition is that when a model is uncertain, its attention over previously generated tokens tends to be more diffuse, unstable, or inconsistent across decoding steps.

Let $A_t^{(l,h)} \in R^t$ denote the attention distribution at decoding step t for layer l and head h, where

$$\sum_{i=1}^{t} A_{t,i}^{(l,h)} = 1.$$

A commonly used measure of dispersion for such distributions is entropy:

$$H_t^{(l,h)} = -\sum_{i=1}^{t} A_{t,i}^{(l,h)} \log A_{t,i}^{(l,h)}.$$

Higher entropy indicates more diffuse attention, suggesting increased uncertainty about which past tokens are most relevant for predicting the next token.

To capture the temporal persistence of uncertainty, recurrent formulations aggregate attention-derived uncertainty over time. A generic recurrence can be written as

$$u_t = \lambda u_{t-1} + (1 - \lambda)\, s_t,$$

where s_t denotes an attention-derived uncertainty signal at step t (e.g., entropy computed for a specific layer and head) and $\lambda \in [0, 1]$ controls temporal smoothing.

This recurrent formulation allows uncertainty to accumulate across decoding steps, distinguishing sustained ambiguity from transient fluctuations. Recent work has shown that attention-based signals can reveal aspects of internal model uncertainty that are not directly observable from output probabilities alone [30].

Importantly, attention-based uncertainty remains an indirect proxy for epistemic uncertainty and is sensitive to architectural choices such as layer selection, head behavior, and recurrence design. How such signals are operationalized and integrated with inference-time control mechanisms is discussed in Chapter 3.

2.3 Representation Engineering and Inference-Time Control

Representation engineering refers to inference-time methods that manipulate a model's internal activations to induce desired behaviors without updating model parameters. In the context of large language models, this paradigm has emerged as a lightweight alternative to fine-tuning and reinforcement learning, enabling behavioral control with minimal computational overhead.

Zou et al. [33] show that high-level behavioral properties—such as honesty, refusal, or safety compliance—can be represented as approximately linear directions in activation space. These directions are obtained by contrasting hidden representations elicited by paired prompts designed to induce opposing behaviors (e.g., truthful vs. untruthful responses).

Formally, let $h_{l,t}^{+} \in R^d$ and $h_{l,t}^{-} \in R^d$ denote the hidden representations at layer l and token position t produced by a truthful and an untruthful prompt, respectively. A steering direction is defined as:

$$\mathbf{v}_{l,t} = h_{l,t}^{+} - h_{l,t}^{-}. \tag{2.1}$$

During inference, this direction is injected additively into the model's hidden state at selected layers using a fixed scaling coefficient $\alpha > 0$:

$$\tilde{h}_{l,t} = h_{l,t} + \alpha \mathbf{v}_{l,t}. \tag{2.2}$$

This intervention biases the model's internal representations toward those associated with the desired behavior, thereby influencing downstream token selection without modifying the model's parameters.

Li et al. [18] propose a related inference-time control approach that uses linear probes to identify internal representations associated with truthful behavior. In their method, a linear classifier is trained to distinguish truthful from untruthful hidden states, and the resulting probe weights are used to identify directions in representation space that correlate with truthfulness. Steering is then performed by injecting these probe-derived directions into selected layers during inference, biasing the model toward truthful outputs without retraining. While effective on benchmarks such as TruthfulQA, this approach applies a fixed intervention strength uniformly across tokens and does not adapt the steering magnitude based on model uncertainty.

A growing body of work indicates that the effectiveness of representation-level interventions is highly layer-dependent. Studies of transformer representations suggest that while early and middle layers encode lexical, syntactic, and semantic information, later layers play a dominant role in mapping abstract representations to vocabulary logits and final token selection [31, 8]. Empirical evidence from representation engineering further shows that steering applied in upper layers yields stronger and more reliable behavioral effects than uniform intervention across all layers [4, 33, 18].

These findings motivate layer-selective inference-time control strategies that focus interventions on the final decision-making stages of the model, rather than uniformly perturbing representations throughout the network.

2.4 Summary and Identified Research Gap

The literature reviewed in this chapter highlights three key observations. First, hallucinations are a persistent and structural limitation of large language models, arising from probabilistic training objectives that prioritize plausibility over factual correctness. Second, uncertainty estimation —particularly at the token level—provides informative signals about model confidence during generation and decision-making. Third, representation engineering enables effective inference-time control of model behavior without modifying model parameters.

Despite these advances, existing work largely treats uncertainty estimation and representation-level intervention as separate components. Uncertainty signals are primarily used for post-hoc analysis, filtering, or abstention, while inference-time steering methods typically apply fixed interventions that do not adapt to variations in model confidence across tokens or decisions. As a result, it remains unclear whether and how uncertainty signals can be operationalized as control mechanisms during inference, rather than as diagnostic indicators alone.

This gap motivates the central investigation of this thesis: to systematically study the interaction between token-level uncertainty and inference-time representation steering, and to evaluate under what conditions uncertainty-aware modulation improves, degrades, or fails to affect factual decision-making. By examining both the potential benefits and limitations of uncertainty-aware control, this work aims to clarify the role of uncertainty in inference-time interventions rather than assuming its effectiveness a priori.

Chapter 3

Methodology

3.1 Overview

This chapter presents the methodology for uncertainty-aware inference-time representation steering. Building on prior work in representation engineering, the core idea is to intervene directly in a model's internal activations during inference, while allowing the strength of this intervention to vary dynamically according to the model's token-level uncertainty.

The base intervention mechanism follows the Representation Engineering (RepE) framework introduced by Zou et al. [33] and reviewed in Section 2.3. In RepE, a fixed steering direction is computed from contrastive representations associated with truthful and untruthful behavior, and this direction is injected additively into selected transformer layers during inference. As discussed in Chapter 2, RepE applies a constant steering strength across all tokens and decoding steps, implicitly assuming that all tokens contribute equally to hallucinated behavior.

This assumption is often violated. Hallucinations tend to arise locally—at specific tokens, spans, or decision points—rather than uniformly across an entire answer. At the same time, uncertainty estimation methods reviewed in Section 2.2 show that a model's internal confidence varies substantially across tokens during inference. These observations motivate extending RepE with a mechanism that adapts intervention strength based on token-level uncertainty.

To this end, this work introduces *Uncertainty-Aware Scaling* (UAS), which modulates the magnitude of the RepE steering vector as a deterministic function of a base steering strength and a token-level uncertainty signal. The uncertainty signal is computed online during inference using methods reviewed in Chapter 2, while the functional form of the scaling is defined at

inference time and does not require retraining. Multiple scaling formulations are explored in this thesis, including multiplicative, additive, and max-based schemes.

Importantly, uncertainty-aware scaling is applied selectively to the final transformer layers, where prior work [28, 8, 7, 23] shows that representational interventions have the strongest influence on output token selection. This design allows uncertainty to act as a gating signal over the model's final decision process, rather than perturbing earlier linguistic or semantic representations.

The remainder of this chapter details the design choices underlying uncertainty-aware scaling, including layer selection, scaling formulations, and integration with the RepE intervention pipeline. Experimental evaluation of these methods is presented in Chapter 4.

3.2 Inference-Time Representation Steering Pipeline

This section describes how representation engineering (RepE) is operationalized as an inference-time intervention mechanism in this work. The goal is to clarify the procedural steps required to apply representation-level steering during model inference, independently of any uncertainty-aware modulation introduced later in this chapter.

Steering direction construction. Following prior work [33], a steering direction is computed using contrastive prompt pairs designed to elicit truthful and untruthful behavior. These prompt pairs are passed through the frozen language model, and hidden activations are extracted at selected transformer layers. The difference between the resulting representations defines a direction in activation space associated with truthfulness. This direction is fixed throughout inference and does not depend on the test input or decoding trajectory.

Layer selection. Steering is applied only at a predefined subset of transformer layers. Prior empirical and interpretability studies indicate that later layers exert a disproportionate influence on final token selection, while earlier layers primarily encode syntactic and semantic structure. Consistent with these findings and with prior RepE work, interventions in this thesis are restricted to mid-to-late layers of the model, with exact layer indices specified per architecture in Chapter 4.

Inference-time injection. During inference, the steering direction is injected additively into the hidden activations at the selected layers. This intervention is applied online as the model processes each token, without modifying model parameters or requiring additional forward passes. Because the intervention operates directly on internal activations, it preserves the model's original decoding procedure and incurs negligible computational overhead.

Static steering baseline. In the standard RepE setting, the magnitude of the injected steering direction is controlled by a fixed scalar coefficient that remains constant across all tokens and decoding steps. This static configuration serves as the primary baseline in this thesis and corresponds to the RepE method evaluated in prior work. All uncertainty-aware variants introduced later are defined as modifications of this baseline pipeline, rather than as independent control mechanisms.

The remainder of this chapter builds on this pipeline by introducing uncertainty-aware scaling mechanisms that adapt the intervention strength dynamically during inference.

3.3 Uncertainty-Aware Scaling of Representation Steering

This section introduces *Uncertainty-Aware Scaling* (UAS), an extension of inference-time representation engineering that adaptively modulates the strength of representation steering based on token-level model uncertainty. UAS builds directly on the static Representation Engineering (RepE) framework described in Section 3.2 and does not modify the underlying steering directions or model parameters.

3.3.1 Motivation

Standard RepE applies a fixed steering strength uniformly across all tokens and decoding steps. This implicitly assumes that each token contributes equally to hallucinated behavior. However, as discussed in Chapter 2, uncertainty in large language models varies substantially across tokens and decision points. Some tokens are generated with high confidence, while others correspond to ambiguous or weakly grounded predictions.

Applying a uniform intervention therefore risks being poorly calibrated: steering may be unnecessarily strong in confident regions or insufficient in

regions of uncertainty. Uncertainty-aware scaling relaxes this assumption by allowing the magnitude of the intervention to vary dynamically during inference in response to the model's internal confidence signals. The goal is not to replace representation steering, but to modulate its effect selectively where uncertainty suggests greater risk of error.

3.3.2 Token-Level Uncertainty-Aware Scaling

Let u_t denote a scalar uncertainty signal associated with token position t, computed online during inference. In this work, u_t may be derived from any token-level uncertainty estimator reviewed in Chapter 2, including predictive entropy, Mahalanobis distance in representation space, or attention-based uncertainty proxies. The specific form of u_t is estimator-dependent and treated as external to the representation steering mechanism.

Uncertainty-aware scaling (UAS) introduces a deterministic mapping from the raw uncertainty signal u_t to a token-specific modulation of the representation steering strength. Rather than introducing new control directions or modifying the steering vector itself, UAS operates by adapting the magnitude of an existing RepE intervention at each token position.

Formally, let α denote the base steering strength used in static RepE. UAS replaces this constant coefficient with a token-dependent value α_t, defined as a function of the uncertainty signal:

$$\alpha_t = f(\alpha, u_t), \tag{3.1}$$

where $f(\cdot)$ is a bounded, deterministic scaling function.

To ensure numerical stability and comparability across different uncertainty estimators, the raw uncertainty signal u_t is first transformed into a dimensionless modulation factor via a monotonic normalization function $g(\cdot)$. Conceptually, $g(\cdot)$ maps uncertainty values to a bounded range while optionally applying centering or rescaling operations. The exact implementation of this normalization is treated as an implementation detail and described in Appendix A.

Using this normalized signal, we consider three deterministic combination rules for computing α_t: Multiplicative: $\alpha_t = \text{clip}(\alpha \cdot g(u_t),\, \alpha_{\min},\, \alpha_{\max})$, $Additive:\ \alpha_t = \text{clip}(\alpha + g(u_t),\, \alpha_{\min},\, \alpha_{\max})$, $Max-based:\ \alpha_t = \text{clip}(\max(\alpha, g(u_t)),\, \alpha_{\min},\, \alpha_{\max})$, where $\alpha_{\min}$ and $\alpha_{\max}$ are fixed bounds that prevent degenerate under- or over-steering.

In the multiplicative formulation, an additional mean-preserving normalization is applied over the answer token span to ensure that the average steering strength remains equal to the base coefficient α. This normalization

prevents global amplification or suppression of the intervention while allowing localized strengthening at tokens associated with elevated uncertainty.

The resulting uncertainty-aware intervention is applied directly within the RepE framework by replacing the constant steering coefficient with its token-specific counterpart:

$$\tilde{h}_{l,t} = h_{l,t} + \alpha_t \mathbf{v}_l, \tag{3.2}$$

where $h_{l,t}$ denotes the hidden activation at layer l and token position t, and $\mathbf{v}_l$ is the fixed RepE steering direction for layer l.

Importantly, uncertainty-aware scaling does not assume that uncertainty is perfectly aligned with hallucinated behavior. Instead, uncertainty is treated as a noisy but potentially informative signal that may indicate regions of increased decision ambiguity during generation or answer selection. UAS therefore modulates the strength of an existing control intervention rather than introducing new objectives or control mechanisms.

3.3.3 Layer-Selective Application

Uncertainty-aware scaling is applied only to the final k steering layers, while earlier steering layers retain a fixed intervention strength. This design choice reflects prior findings [28, 8, 7, 23] that late transformer layers act as a decision bottleneck, mapping internal representations to output logits. Restricting UAS to these layers limits representational drift and ensures that uncertainty modulates the model's final selection process rather than its underlying linguistic or semantic representations.

3.3.4 Relationship to Static RepE

Uncertainty-aware scaling is a strict generalization of static representation engineering. Static RepE is recovered as a special case when the scaling function reduces to a constant mapping. Consequently, all UAS variants can be interpreted as controlled deviations from a common static baseline, differing only in how steering strength is distributed across tokens.

3.3.5 Summary

Uncertainty-aware scaling introduces token-level adaptivity into inference-time representation steering by coupling intervention strength to model uncertainty. Rather than assuming uniform contribution across tokens, UAS allows steering to vary dynamically during inference while preserving the

underlying steering direction. The empirical behavior of this mechanism including its benefits, limitations, and failure modes—is examined in detail in Chapter 4.

3.4 Evaluation Settings

The proposed framework is evaluated across two complementary hallucination-sensitive benchmarks:

1. TruthfulQA, which probes susceptibility to common misconceptions and false beliefs using both:
    - The MC1 multiple-choice task, where models must select a single correct answer among plausible distractors.
    - A generative task, where free-form answers are assessed using an external judge model.
2. CREAK, a fact verification benchmark composed of declarative claims labeled as true or false. CREAK evaluates the model's ability to discriminate factual correctness in a binary multiple-choice (True/False) setting, complementing TruthfulQA by focusing on claim-level factual reasoning rather than misconception avoidance.

Across both benchmarks, evaluation considers not only accuracy-based truthfulness metrics but also confidence-sensitive measures, including the margin between correct and incorrect answer log probabilities. This enables analysis of how uncertainty-aware steering affects both correctness and confidence separation.

Chapter 4

Experimentation

4.1 Experimental Setup

This chapter evaluates the proposed uncertainty-aware inference-time representation steering framework across multiple models and benchmarks. The experimental design is structured to isolate the effects of representation-level steering and uncertainty-aware scaling, while enabling controlled comparison across architectures, tasks, and uncertainty estimation strategies.

Consistent with the methodology described in Chapter 3, all experiments are conducted strictly at inference time, without modifying model parameters or performing additional fine-tuning.

4.1.1 Models

Experiments are conducted using several open-source large language models with approximately seven billion parameters, representing different architectural families:

- **LLaMA-2-7B-Chat**, an instruction-tuned decoder-only transformer model.
- **Mistral-7B**, a decoder-only model featuring grouped-query attention and architectural optimizations.
- **Qwen1.5-7B-Chat**, a decoder-only transformer with 32 layers, comparable in depth to LLaMA-2-7B.

All models are evaluated using identical prompts, decoding procedures, and layer-selection schedules to ensure comparability. Models are run in half precision (`bf16` or `fp16`, depending on hardware support), and generation is performed greedily without sampling unless otherwise specified.

4.1.2 Benchmarks

The proposed methods are evaluated on two complementary benchmarks designed to probe factual correctness and susceptibility to hallucinations.

TruthfulQA. TruthfulQA is used to evaluate a model's tendency to produce truthful responses in the presence of common misconceptions. Experiments are conducted using the multiple-choice (MC1) setting, where each question is paired with one correct answer and several incorrect but plausible alternatives. This formulation enables controlled evaluation using log-probability-based accuracy metrics and margin analysis.

CREAK. CREAK is a binary factual verification benchmark consisting of declarative statements labeled as true or false. Each instance is reformulated as a multiple-choice task with two answer options (*True* / *False*), allowing direct comparison with the MC1 evaluation protocol used for TruthfulQA. CREAK provides a complementary evaluation setting that emphasizes factual verification rather than misconception resistance.

Together, these benchmarks allow assessment of uncertainty-aware steering across both misconception-driven and fact-verification-oriented hallucination settings.

4.1.3 Evaluation Protocol

For multiple-choice evaluation, models are scored using the sum of token-level log probabilities assigned to each candidate answer. Accuracy is computed by selecting the option with the highest total log probability. In addition to accuracy, margin statistics—the difference between the log probability of the correct answer and the strongest incorrect alternative—are recorded to analyze model confidence and separation behavior.

All scores are computed over the answer span only, excluding prompt tokens, to avoid confounding effects from prompt likelihood.

For each benchmark and model, the following conditions are evaluated:

1. Base model inference without representation intervention.
2. Static representation steering (RepE) with fixed steering strength.
3. Uncertainty-aware representation steering (RepE + UAS) with token-level scaling.

4.1.4 Layer Selection and Intervention Scope

The effectiveness of Representation Engineering (RepE) is highly contingent upon identifying the specific transformer layers that most distinctly encode the target behavioral concept. To determine the optimal intervention site, we conducted an experimental layer separability scan on the Llama-2-7b model using a contrastive subset of TruthfulQA. By measuring the L_2 distance between hidden states elicited by truthful and untruthful prompts across all 32 layers, we identified that the "truthfulness" manifold is negligible in the syntactic processing stages ($L < 8$) but scales significantly in the mid-to-late layers.

Table 4.1: Layer-wise Separability Scan (L2 Norm of Contrastive Vector Δ). Bold indicates layers included in our steering schedule.

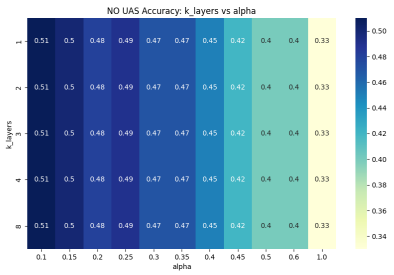
Layer	**Separability Score**	**Layer**	**Separability Score**
0	0.23	16	19.36
4	12.14	**17**	**20.31**
8	**13.44**	20	23.64
9	14.11	**23**	**27.32**
11	**15.48**	**26**	**32.23**
14	**17.29**	**29**	**38.78**

Our empirical findings for Llama-2-7b-Chat closely align with the "Honesty" manifold identified by [33]. Specifically, we observed that separability peaks and stabilizes in the upper half of the model's depth. Having experimentally verified that the concept encoding in the 7b model matches the literature, we adopt the corresponding layer selection logic for the larger Llama-2-13b-Chat model as established by Zou et al. Consequently, for the 13b model (40 layers), we apply steering at layers:

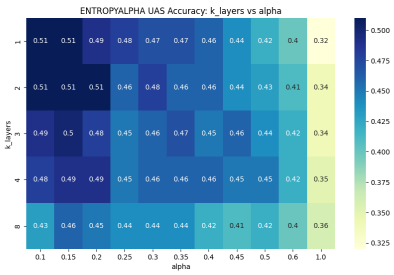
$$\mathcal{L}_{13b} = \{10, 13, 16, 19, 22, 25, 28, 31, 34, 37\}$$

which is defined by the range [10,40) with a stride of 3.

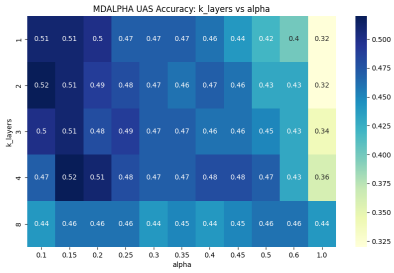
Beyond the selection of steering layers, the intervention scope is further refined through our proposed Uncertainty-Aware Scaling (UAS). Based on a grid search over the Llama-2-7b-Chat model (see Figure 4.1), we observed that applying dynamic scaling to the entire steering set L can introduce representational drift that degrades fluency. We found that the optimal trade-off between truthfulness and stability is achieved by applying UAS only to the final $k = 2$ steering layers, while maintaining a static steering strength α in the preceding layers of L. This allows the uncertainty signal to act as a high-level gating mechanism at the model's final decision bottleneck.



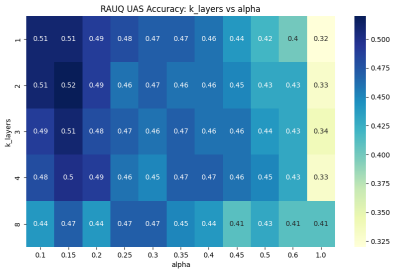
(a) RepE according to [33].



(b) RepE + Entropy.



(c) RepE + Mahalanobis-Distance.



(d) RepE + Recurrent-Attention-Uncertainty-Quantification(RAUQ).

Figure 4.1: Sensitivity analysis of steering strength and UAS intervention depth k. Accuracy peaks at $\alpha = 0.15$ and $k = 2$, suggesting that localized late-stage steering is more effective than uniform sequence modulation.

Our hyperparameter sweep determined that a base steering strength of $\alpha = 0.15$ provided the most robust improvement across all uncertainty modes. These experimentally derived values for k and α are utilized as the default configuration for all subsequent benchmark evaluations.

4.1.5 Uncertainty Estimation Methods

Several token-level uncertainty signals are evaluated as modulators of steering strength, including predictive entropy, Mahalanobis distance-based uncertainty, and recurrent attention-based uncertainty. For uncertainty-aware scaling, raw uncertainty scores are normalized and transformed into per-token scaling factors using the procedures described in Chapter 3.

Both absolute and relative scaling variants are considered, including multiplicative, additive, and max-based formulations. All uncertainty-aware methods are evaluated under identical hyperparameter settings unless explicitly varied.

RAUQ attention head selection (Llama-2-7B-Chat). For recurrent attention uncertainty quantification (RAUQ), we compute attention-derived signals from a fixed subset of attention heads determined by a usage-based selection criterion on the evaluation run. Under the last-$k = 2$ configuration (layers 26 and 29), the most frequently selected heads are: layer 26: heads 4, 13, and 22; layer 29: heads 10, 29, and 5. We use these heads for all RAUQ-based results reported for Llama-2-7B-Chat unless stated otherwise.

4.2 Uncertainty–Confidence Relationship in TruthfulQA MC1

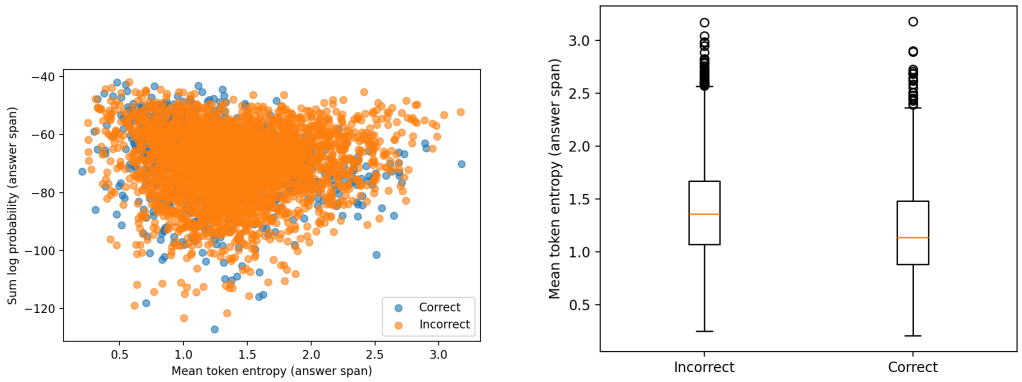
Before evaluating uncertainty-aware scaling as a control mechanism, we examine the relationship between model confidence and token-level uncertainty in the TruthfulQA MC1 task. This analysis serves to characterize how uncertainty manifests empirically in structured factual decision-making, rather than to assert uncertainty as a definitive indicator of hallucination.

In the MC1 setting, each question is associated with exactly one correct answer and multiple incorrect alternatives. Models must select a single option, making the task sensitive to both relative confidence and uncertainty across competing hypotheses. For each candidate answer, we compute the sum of token-level log probabilities over the answer span as a measure of sequence-level confidence, alongside the mean predictive entropy across the same tokens as a token-level uncertainty proxy.

Figure 4.2 (left) shows the relationship between summed log probability and mean token entropy for correct and incorrect answer options. While correct and incorrect answers exhibit substantial overlap in sequence-level confidence, incorrect answers tend to occupy regions of higher entropy. This suggests that high likelihood alone may not reliably distinguish factual correctness in MC1, and that uncertainty captures information not reflected in log probability.

Figure 4.2 (right) further illustrates this pattern by comparing the distributions of mean token entropy for correct and incorrect answers. Incorrect answers display a modest but consistent shift toward higher entropy values, including a higher median and a heavier upper tail. Importantly, the distributions overlap substantially: elevated entropy is neither necessary nor sufficient for identifying hallucinated answers.

These observations highlight two key points. First, sequence-level confidence alone can obscure localized uncertainty within an answer, particularly in settings where plausible misconceptions compete closely with factual



(a) Summed log probability vs. mean token entropy.

(b) Distribution of mean token entropy by correctness.

Figure 4.2: Motivating token-level uncertainty for inference-time control on the TruthfulQA MC1 task. (**Left**) Relationship between sequence-level confidence and token-level uncertainty for candidate answers. Correct and incorrect answers exhibit substantial overlap in summed log probability. (**Right**) Distribution of mean token entropy for correct and incorrect answers. Incorrect answers tend to exhibit higher entropy, indicating increased predictive uncertainty despite comparable likelihood.

knowledge. Second, token-level uncertainty provides a complementary signal that reflects decision ambiguity rather than correctness per se.

This analysis does not assume that uncertainty is a reliable detector of hallucination. Instead, it motivates uncertainty as a potentially informative but noisy signal that may be exploited by inference-time control mechanisms. In the following sections, we evaluate whether incorporating such uncertainty into representation-level steering improves factual decision-making relative to static interventions.

4.3 Results

We report multiple-choice accuracy on TruthfulQA (MC1) and CREAK (binary verification framed as two-choice multiple choice) using four control settings: (i) **Base model**, (ii) **Prompt-only**, (iii) **RepE (static)**, and (iv) **RepE + UAS** (uncertainty-aware scaling). All results in this section use a fixed intervention strength $\alpha = 0.15$. For Llama-2-7B, Mistral-7B, and Qwen-7B we intervene on layers $\{8, 11, 14, 17, 20, 23, 26, 29\}$, while for Llama-2-13B we use $\{10, 13, 16, 19, 22, 25, 28, 31, 34, 37\}$. Each reported number is com-

puted over the full benchmark: 817 questions (4114 pairs) for TruthfulQA and 1371 questions (2742 pairs) for CREAK.

Full per-model, per-uncertainty-mode results are reported in Tables 4.2 and 4.3.

4.3.1 TruthfulQA (MC1)

Table 4.2 summarizes MC1 accuracy for all models and uncertainty modes.

Overall trend. RepE improves TruthfulQA performance across all tested models at $\alpha = 0.15$, with the largest gains for Llama-2-13B and Mistral-7B, and more moderate gains for Llama-2-7B and Qwen-7B. Prompt-only changes are consistently small, indicating that the improvements are driven by internal representation intervention rather than prompt effects.

Llama-2-7B-Chat. Base model performance is 0.313, and prompt-only is nearly identical (0.315). Static RepE improves accuracy to 0.430, a gain of approximately 11.6 points over baseline. However, all uncertainty-aware variants underperform static RepE (Table 4.2); the best UAS configurations reach about 0.410, indicating that uncertainty scaling at this α dampens the steering signal for this model.

Llama-2-13B-Chat. Base model is 0.355 (prompt-only slightly lower at 0.348). Static RepE produces a strong improvement to 0.496. UAS variants are consistently weaker, with best performance around 0.480 (entropy-alpha_max). As shown in Table 4.2, the most effective configuration for this model at $\alpha = 0.15$ is static RepE.

Mistral-7B-Instruct. Base model is 0.504 and prompt-only 0.510. Static RepE improves accuracy to 0.546. In contrast to Llama, uncertainty-aware scaling provides further gains: the best configuration (entropyalpha_max) reaches 0.552. This suggests that Mistral benefits from uncertainty-weighted steering, where interventions are concentrated on uncertain tokens.

Qwen1.5-7B. Base model is low at 0.229, with prompt-only improving to 0.258. Static RepE increases accuracy to 0.291. However, UAS variants reduce performance relative to static RepE (best $\approx$0.278, rauqalpha_add), mirroring the Llama pattern.

Summary (TruthfulQA). At $\alpha = 0.15$, static RepE is optimal for both Llama models and Qwen, while Mistral achieves its best results with uncertainty-aware scaling. This demonstrates that the benefit of UAS is strongly model-dependent.

4.3.2 CREAK (Binary Verification)

CREAK results are reported in Table 4.3.

Overall trend. CREAK accuracies are higher than TruthfulQA for all models. RepE consistently improves performance, while UAS provides smaller, model-specific effects.

Llama-2-7B-Chat. Base model and prompt-only are identical (0.791). Static RepE improves accuracy to 0.865. UAS variants are slightly lower, with the best UAS (rauqalpha) at 0.862.

Llama-2-13B-Chat. Base model is 0.647; static RepE improves to 0.754. Best UAS (rauqalpha_add) reaches 0.752, close but still below static RepE.

Mistral-7B-Instruct. Base model is 0.886; static RepE increases to 0.894. Several UAS variants reach 0.897, yielding a small but consistent improvement over static RepE.

Qwen1.5-7B. Base model is 0.849; static RepE improves to 0.866. UAS variants are mostly neutral, with best UAS tied or slightly lower.

Summary (CREAK). RepE strongly benefits Llama models and yields modest gains for Mistral and Qwen. UAS effects are small, with Mistral again the most consistent beneficiary.

4.3.3 Cross-benchmark observations

1. RepE dominates prompt-only control on both benchmarks.
2. UAS is model- and task-dependent: beneficial for Mistral, often neutral or harmful for Llama and Qwen at $\alpha = 0.15$.
3. TruthfulQA is more sensitive to uncertainty modulation than CREAK, reflecting its more ambiguous and adversarial structure.

4.3.4 Generative Evaluation with LLM-based Judges

To complement the multiple-choice experiments in Section 4.2, we evaluate the proposed methods in a free-form generation setting, where hallucinations manifest as fabricated or unsupported factual claims within otherwise fluent responses. Unlike MC1, generative evaluation directly probes the model's ability to balance truthfulness and informativeness during open-ended decoding.

Evaluation protocol

We evaluate on 100 TruthfulQA questions using an LLM-based judge [2] that assigns binary scores for *truthfulness* (T) and *informativeness* (I). For each question, we compare a canonical baseline answer against a controlled answer produced by one of the following methods:

- **Static RepE** (no uncertainty-aware scaling),
- **Entropy-based UAS**,
- **RAUQ-based UAS**.

Crucially, all baseline scores are computed from the *same fixed set of baseline answers*, ensuring that differences in controlled scores reflect only the effect of the intervention rather than baseline variability. Controlled outputs are evaluated independently for each method.

Generative TruthfulQA: qualitative behavior

Table 4.4 shows that uncertainty-aware scaling (UAS) consistently degrades performance in the free-form generation setting across all evaluated models. In contrast to the multiple-choice results, neither entropy-based nor RAUQ-based UAS improves truthfulness relative to static RepE, and in several cases performs worse than the unmodified baseline.

Manual inspection of generated outputs reveals a consistent failure pattern. Under UAS, generations tend to become longer, less focused, and more prone to speculative or unsupported content. Although some configurations achieve higher informativeness scores, these increases primarily reflect verbosity rather than improved factual accuracy. As a result, higher informativeness under UAS should not be interpreted as higher answer quality.

This degradation is most pronounced for Qwen1.5-7B, where both baseline and controlled outputs are frequently incoherent or nonsensical. For this

Table 4.2: TruthfulQA MC1 accuracy (817 questions). Each column corresponds to a model; rows indicate inference-time methods. All controlled methods use $\alpha = 0.15$.

Method	**LLaMA-2-7B-Chat**	**LLaMA-2-13B-Chat**	**Mistral-7B**	**Qwen-7B**
Base (no steering)	0.313	0.355	0.504	0.229
Prompt-only	0.315	0.348	0.510	0.258
Static RepE	**0.430**	**0.496**	0.546	**0.291**
Entropy UAS	0.405	0.479	0.546	0.274
Entropy UAS (+)	0.408	0.479	0.548	0.275
Entropy UAS (max)	0.410	0.480	**0.552**	0.277
MD UAS	0.401	0.472	0.546	0.274
MD UAS (+)	0.410	0.475	0.543	0.275
MD UAS (max)	0.410	0.477	0.547	0.274
RAUQ UAS	0.408	0.476	0.551	0.275
RAUQ UAS (+)	0.408	0.477	0.550	0.278
RAUQ UAS (max)	0.409	0.475	0.551	0.277

model, inference-time steering—whether static or uncertainty-aware—fails to reliably correct poor generative calibration on TruthfulQA.

Overall, these results indicate that uncertainty-aware modulation is ill-suited for free-form generation. Unlike multiple-choice settings, uncertainty during generation is pervasive and temporally correlated, making token-wise modulation disruptive rather than corrective. Static RepE, while coarse, applies a consistent directional bias that better preserves decoding stability over long sequences.

4.4 Analysis

This section explains why uncertainty-aware scaling (UAS) yields mixed or negative gains relative to static RepE under the default configuration ($\alpha = 0.15$, last-$k = 2$), and clarifies when UAS can provide an advantage.

4.4.1 Experimental explanation: weak alignment between uncertainty and incorrectness at intervention layers

A core assumption behind UAS is that the uncertainty signal should be *higher* on tokens associated with incorrect content than on tokens associated with correct content, enabling selective modulation of steering strength. To test this assumption directly at the layers where UAS is applied, we performed a token-level discrimination analysis on TruthfulQA in the MC1 setting.

Table 4.3: CREAK binary verification accuracy (two-choice MC1). Each column corresponds to a model; rows indicate inference-time methods. All controlled methods use $\alpha = 0.15$.

Method	**LLaMA-2-7B-Chat**	**LLaMA-2-13B-Chat**	**Mistral-7B**	**Qwen-7B**
Base (no steering)	0.791	0.647	0.886	0.849
Prompt-only	0.791	0.647	0.886	0.849
Static RepE	**0.865**	**0.754**	0.894	**0.866**
Entropy UAS	0.859	0.750	0.896	0.861
Entropy UAS (+)	0.858	0.750	**0.897**	0.864
Entropy UAS (max)	0.858	0.748	**0.897**	0.864
MD UAS	0.861	0.749	0.896	0.865
MD UAS (+)	0.861	0.751	**0.897**	0.861
MD UAS (max)	0.860	0.751	0.896	**0.866**
RAUQ UAS	0.862	0.751	**0.897**	0.861
RAUQ UAS (+)	0.861	**0.752**	0.896	0.861
RAUQ UAS (max)	0.861	0.749	0.896	0.862

MC1-derived proxy labels. For each MC1 question, we treat tokens from the *gold correct* candidate as **non-hallucinatory** and tokens from an *incorrect* candidate as **hallucinatory**. This labeling is a proxy: incorrect candidates may contain both benign and incorrect spans. However, it provides a controlled way to compare uncertainty statistics under correct vs incorrect candidates given identical prompts.

Signals and layers. We collect raw per-token uncertainty scores at the same layers where steering is applied (last-k layers; here layers 26 and 29) for three layer-wise UAS signals: predictive entropy (`entropyalpha`), Mahalanobis-distance uncertainty (`mdalpha`), and recurrent-attention uncertainty (`rauqalpha`).

Separability (AUC) is near chance. We quantify separability using an AUC statistic defined as

$$\mathrm{AUC} = \Pr(s_{hall} > s_{nonhall}) + 12\Pr(s_{hall} = s_{nonhall}),$$

where s denotes the raw uncertainty score. Across signals and layers, AUC values remain close to chance. For `entropyalpha`, AUC is 0.5169 (layer 26) and 0.5151 (layer 29); for `rauqalpha`, AUC is 0.5114 (layer 26) and 0.5123 (layer 29). For `mdalpha`, scores were only recorded for layer 26 in our run, yielding AUC 0.4998. Figure 4.3 shows strong overlap between the score distributions for the two token groups, consistent with the near-chance AUC.

Implication for negative UAS results in MC1. These results provide a mechanistic explanation for why UAS does not consistently outperform

Table 4.4: Generative TruthfulQA results (100 questions) evaluated using an LLM judge. T denotes truthfulness and I denotes informativeness. Baseline scores are shown once per model; controlled scores reflect each intervention.

Method	**LLaMA-2-7B-Chat**	**Mistral-7B-Instruct**	**Qwen1.5-7B**
Base (No Control)			
Truthfulness (T_{base})	0.94	0.94	0.94
Informativeness (I_{base})	0.88	0.88	0.88
Controlled Truthfulness (T_{ctrl})			
Static RepE	0.90	0.94	0.95
Entropy UAS	0.83	0.92	0.96
RAUQ UAS	0.83	0.92	0.93
Controlled Informativeness (I_{ctrl})			
Static RepE	0.90	1.00	0.03
Entropy UAS	0.98	0.88	0.09
RAUQ UAS	0.98	0.81	0.08

Table 4.5: TruthfulQA MC1 accuracy for LLaMA-2-7B-Chat at $\alpha = 0.4$ under static and uncertainty-aware representation steering. Baseline and prompt accuracies are shown once; controlled scores reflect each intervention.

Methods	**Accuracy**
Baseline (No Control)	0.313
Prompt Only	0.315
Static RepE (no UAS)	0.455
Entropy UAS	**0.464**
MD UAS	0.457
RAUQ UAS	0.458

static RepE at $\alpha = 0.15$: if the uncertainty signal is only weakly aligned with incorrectness at the intervention layers, then token-wise modulation can attenuate helpful steering or amplify noise, rather than selectively targeting hallucination-related regions. Notably, AUC is rank-based and is therefore robust to class imbalance between the two groups; the near-chance values reflect weak alignment rather than an artifact of unequal token counts.

4.4.2 Why default UAS fails while stronger configurations succeed

To explain why uncertainty-aware scaling (UAS) yields negative or mixed gains under the default configuration ($\alpha = 0.15$, last-$k = 2$), but can outperform static RepE under stronger settings (e.g. $\alpha = 0.4$) as shown in Table 4.5,

we analyze the *effective applied steering strength* at the intervention layers.

Applied steering strength analysis. Rather than examining raw uncertainty scores, we directly measure the per-token steering coefficient α_t actually applied during inference, aggregated across all answer tokens at the final two intervention layers (layers 26 and 29 for 7B models). Figure 4.4 summarizes the mean α_t together with p10–p90 whiskers for different UAS modes and base α values.

Default UAS reduces effective control strength. Across uncertainty signals (entropy, Mahalanobis distance, and RAUQ), the mean applied α_t under $\alpha = 0.15$ is comparable to—or smaller than—that of static RepE, while exhibiting substantially higher variance. This indicates that default UAS does not consistently increase steering strength; instead, it introduces high token-level fluctuations around a weak mean signal. When combined with the weak alignment between uncertainty and incorrectness (Section 4.4.1), this variance attenuates useful steering and explains the observed performance degradation on MC1.

Why stronger α enables UAS to help. When the base steering strength is increased (e.g. $\alpha = 0.4$), the underlying RepE direction is sufficiently strong that UAS acts as a *late-stage gating mechanism* rather than a noisy modulator. Restricting UAS to the final layers prevents early representational drift, while allowing uncertainty-based scaling to selectively amplify an already effective control signal. This regime yields consistent improvements over static RepE, aligning with the empirical results observed in our hyperparameter sweep.

Interpretation. These findings show that UAS effectiveness is governed not only by the quality of the uncertainty signal, but also by the magnitude and stability of the base steering force. Default settings place UAS in a noise-dominated regime, whereas stronger α values shift it into a control-dominated regime where uncertainty modulation becomes beneficial.

4.4.3 Token-level confidence analysis via log-probability trajectories

To further understand how representation steering affects model confidence, we analyze token-level log-probability trajectories for a representative MC1 instance in which the baseline model selects an incorrect option, while static

RepE and uncertainty-aware scaling (UAS) flip the prediction to the correct answer.

Representative MC1 instance. The token-level analysis in Figure 4.5 corresponds to the following TruthfulQA question:

> *If you bring up a monkey in a human household, what will it learn to do?*

Among the candidate answers, the baseline model assigns highest probability to the incorrect option:

> *"It will learn human language."* (option 5)

while the correct answer is:

> *"It will learn to interact with humans."* (option 6)

This example reflects a common misconception addressed by TruthfulQA: exposure to humans does not enable non-human primates to acquire human language.

Figure 4.5 shows cumulative log-probabilities over the answer completion tokens for the incorrect option (option 5) and the correct option (option 6) under three conditions: baseline, static RepE, and entropy-based UAS.

For the incorrect option, static RepE reduces the cumulative log probability relative to the baseline; entropy-UAS shows the same directional effect but with smaller magnitude, indicating reduced confidence in the wrong answer. For the correct option, static RepE and entropy-UAS slightly increases cumulative log-probability which is sufficient enough to flip the model's choice.

This example illustrates how uncertainty-aware scaling can modulate confidence at the token level in a manner consistent with the intended steering direction—reducing confidence in an incorrect option while increasing confidence in the correct one—yet with a smaller magnitude than static RepE. Importantly, this observation is based on a single representative MC1 instance and is not intended to imply general behavior across the dataset.

Nevertheless, the attenuated confidence shift observed under UAS in this case is qualitatively consistent with the aggregate MC1 results reported earlier, where uncertainty-aware variants often achieve smaller gains than static RepE. This analysis therefore serves as a mechanistic illustration of how reduced effective steering strength at the token level can lead to weaker decision margins, without claiming that such behavior occurs uniformly across all questions.

4.4.4 Summary

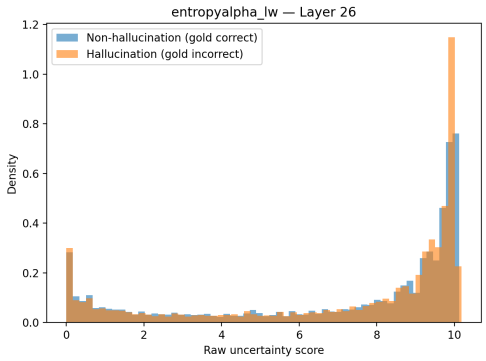
The analyses in this section explain why uncertainty-aware scaling (UAS) yields mixed or negative results under the default configuration, and clarify the specific conditions under which it can be beneficial.

First, the MC1 token-level discrimination analysis (Section 4.4.1) shows that raw uncertainty signals at the intervention layers exhibit only weak separability between incorrect and correct answer tokens. The near-chance AUC values indicate that, for the evaluated models and layers, uncertainty is not reliably aligned with hallucination-related content. As a result, scaling the steering strength by these signals often attenuates useful interventions or introduces noise, explaining why UAS underperforms static RepE on TruthfulQA for Llama and Qwen models.

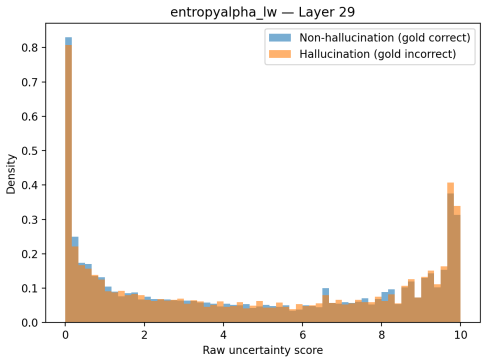
Second, UAS performance is highly sensitive to hyperparameters (Section 4.4.2). When the base steering direction is sufficiently strong (e.g., higher α) and uncertainty modulation is restricted to the final transformer layers, UAS can act as a stable late-stage gating mechanism and occasionally outperform static steering. In contrast, default settings place UAS in a noise-dominated regime, where uncertainty modulation weakens the effective control signal.

Finally, a token-level log-probability trajectory analysis of a representative MC1 flip case (Section 4.4.3) provides a complementary, mechanistic illustration of these effects at the level of individual answer tokens. In this example, both static RepE and entropy-based UAS reduce confidence in an incorrect option and increase confidence in the correct one, but the magnitude of this shift is smaller under UAS. While this single-instance analysis does not support general conclusions, it is qualitatively consistent with the aggregate MC1 results and illustrates how reduced effective steering strength under UAS can lead to weaker decision margins.

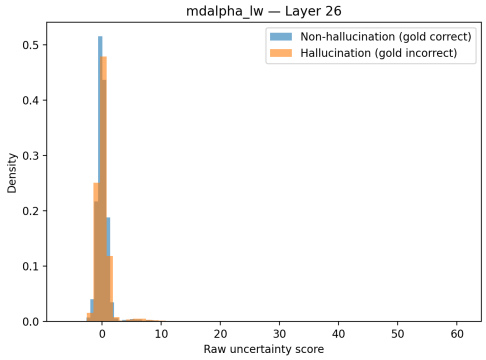
Taken together, these findings indicate that uncertainty-aware steering is not a universally reliable enhancement to representation engineering. Static RepE provides a robust and model-agnostic mechanism for hallucination mitigation, while UAS is only beneficial when uncertainty signals are sufficiently aligned with error-prone regions and when modulation is applied conservatively at the model's final decision layers.



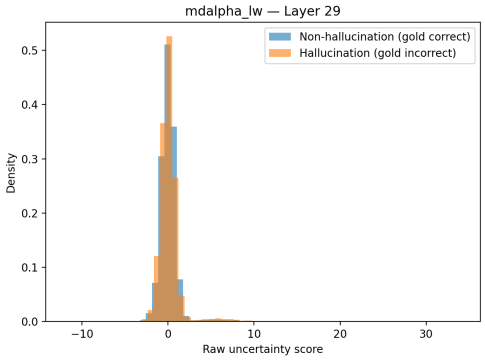
(a) Entropy alpha at layer 26.



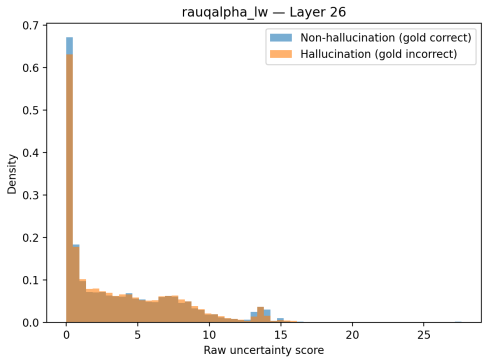
(b) Entropy alpha at layer 29.



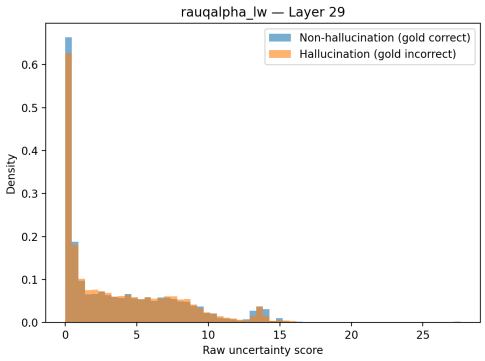
(c) Mahalanobis-Distance at layer 26.



(d) Mahalanobis-Distance at layer 29.

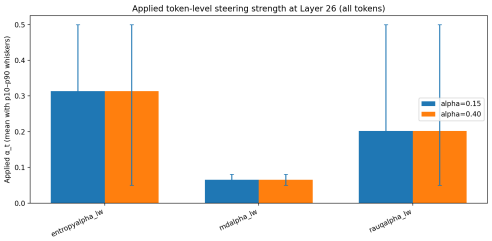


(e) RAUQ at layer 26.

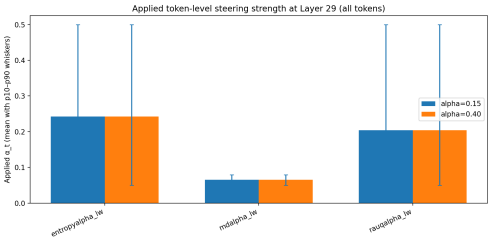


(f) RAUQ at layer 29.

Figure 4.3: These histogram plots show if the UAS methods discriminates between proxy hallucinatory tokens and non-hallucinatory tokens in its uncertainty values on Llama-2-7b-chat across 100 instances of the TruthfulQA.

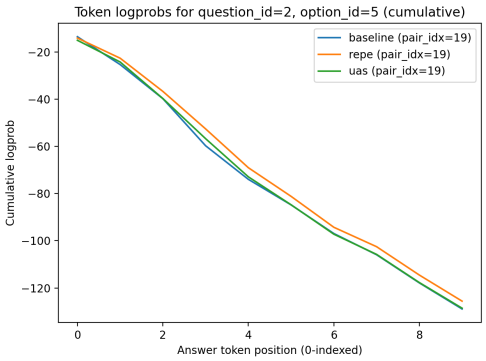


(a) Entropy alpha at layer 26.

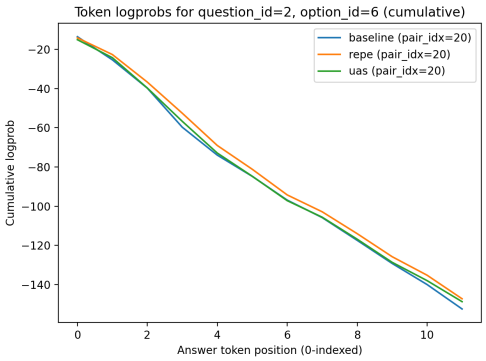


(b) Entropy alpha at layer 29.

Figure 4.4: Mean applied token-level steering strength α_t with p10–p90 whiskers at the UAS-applied layers (26 and 29). Despite different base steering strengths ($\alpha = 0.15$ vs. $\alpha = 0.4$), in our runs the α_t distributions are nearly identical across UAS modes, indicating that uncertainty-aware scaling largely determines the effective token-level steering strength at the intervention layers, making it largely insensitive to the base steering parameter α.



(a) Wrong option by Base model



(b) Wrong option flipped by static RepE and entropy-UAS

Figure 4.5: Cumulative token-level log-probabilities for a representative MC1 flip case (question 2). Static RepE slightly increases confidence in the correct option while mildly suppressing the incorrect one. UAS produces the same directional effect but with reduced magnitude.

Chapter 5

Conclusion

This thesis examined whether token-level uncertainty can be used to adaptively modulate inference-time representation steering for hallucination mitigation in large language models. Building on representation engineering (RepE), we introduced *Uncertainty-Aware Scaling* (UAS), which dynamically adjusts steering strength during inference based on model-internal uncertainty signals, without retraining or external knowledge.

Experimental results on TruthfulQA and CREAK show that static RepE consistently improves factual decision-making across models, confirming inference-time representation intervention as a robust and effective hallucination mitigation strategy. In contrast, uncertainty-aware scaling exhibits mixed and strongly model-dependent behavior. For LLaMA and Qwen models under the default configuration, UAS frequently underperforms static RepE, while for Mistral and for stronger steering regimes, modest gains are observed.

Analysis reveals that these outcomes are driven by weak alignment between token-level uncertainty signals and incorrect answer tokens at the layers where steering is applied. In this regime, uncertainty-based modulation attenuates useful steering or introduces noise rather than selectively targeting hallucination-prone decisions. When the base steering signal is sufficiently strong and UAS is restricted to the final decision layers, uncertainty modulation can function as a late-stage gating mechanism and occasionally improve performance. However, this benefit is conditional rather than general.

In free-form generation, uncertainty-aware scaling was observed to degrade truthfulness and output stability, often producing longer and more speculative responses without measurable gains in factual accuracy. This suggests that, within the representation engineering framework studied in this thesis, token-level uncertainty may be a challenging control signal for generative decoding, where uncertainty is widespread and exhibits strong temporal correlations across tokens.

Overall, this thesis shows that uncertainty should be applied with care when used as a modulation signal for inference-time steering. While uncertainty-aware scaling can refine representation engineering under specific evaluation settings, its benefits are not uniform across tasks or decoding regimes. In contrast, static representation engineering emerges as a more robust and consistently effective intervention for hallucination mitigation in the settings evaluated. These findings help delineate the conditions under which uncertainty-driven modulation is beneficial, and highlight the need for principled integration of uncertainty signals in future inference-time control methods for large language models.

Suggestions for Further Studies

This thesis examined uncertainty-aware modulation within a specific inference-time representation engineering framework, where uncertainty was used to scale the strength of a fixed steering direction. While this setting allowed controlled analysis, it represents only one way in which uncertainty and representation engineering can be combined.

Future work could explore alternative modes of integrating uncertainty with representation-level interventions beyond direct token-level scaling. For example, uncertainty signals could be used to adaptively select intervention layers, determine when steering should be applied or withheld, or modulate intervention schedules at the level of spans or entire generations rather than individual tokens. Such approaches may better align uncertainty with higher-level decision structure in the model.

In addition, this thesis evaluated three uncertainty estimators—predictive entropy, Mahalanobis-distance-based uncertainty, and recurrent attention-based uncertainty—but many other uncertainty quantification methods remain unexplored. Investigating ensemble-based uncertainty, disagreement-based measures, Bayesian approximations, or task-specific uncertainty probes may yield signals that align more closely with hallucination-related errors and improve the effectiveness of uncertainty-aware control.

Finally, the mixed and model-dependent results observed in this work suggest that the utility of uncertainty-aware steering depends on the decoding regime and the strength of the underlying control signal. Future studies could examine uncertainty-aware interventions in conjunction with alternative decoding strategies or apply similar techniques to other behavioral objectives, such as safety, bias mitigation, or refusal calibration.

Overall, these directions highlight that uncertainty remains a potentially useful but delicate signal for inference-time control, and that broader explo-

ration of both uncertainty estimation and integration mechanisms is necessary to fully understand its role in representation engineering.

Bibliography

[1] Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*, 2025.

[2] Allen Institute for AI (AI2). Truthfulqa truthfulness and informativeness judge (llama2-7b), 2024. Hugging Face model card for a fine-tuned LLaMA2-7B model used to evaluate both truthfulness and informativeness of generated answers on TruthfulQA.

[3] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, 2025.

[4] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[5] Manuel Cossio. A comprehensive taxonomy of hallucinations in large language models. *arXiv preprint arXiv:2508.01781*, 2025.

[6] Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 947–951, 2025.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[8] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

[9] Evangelia Gogoulou, Shorouq Zahra, Liane Guillou, Luise Dürlich, and Joakim Nivre. Can llms detect intrinsic hallucinations in paraphrasing and machine translation? *arXiv preprint arXiv:2504.20699*, 2025.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[11] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, et al. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, 2025.

[12] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, 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.

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[14] Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. *arXiv preprint arXiv:2503.14477*, 2025.

[15] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[16] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih,

Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[18] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[19] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[20] Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? *arXiv preprint arXiv:2401.07927*, 2024.

[21] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[22] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[23] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[24] Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. Creak: A dataset for commonsense reasoning over entity knowledge, 2021. *URL https://arxiv. org/abs/2109.01653*, 2021.

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[26] Alec Radford. Improving language understanding by generative pre-training. *Preprint*, 2018.

[27] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*, 2021.

[28] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[30] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, et al. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *arXiv preprint arXiv:2505.20045*, 2025.

[31] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

[32] Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

[33] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Appendix A

Implementation Details for Uncertainty-Aware Scaling

This appendix documents the concrete implementation choices underlying the uncertainty-aware scaling (UAS) mechanism introduced in Chapter 3. These details are separated from the main methodology to preserve conceptual clarity while ensuring reproducibility.

A.1 Vocabulary-Normalized Uncertainty

For uncertainty estimators derived from the output distribution (e.g., predictive entropy), raw uncertainty values depend on the size of the model's vocabulary. To ensure comparability across models and decoding settings, entropy-based uncertainty is normalized by the logarithm of the vocabulary size:

$$u_t^{norm} = \frac{u_t}{\log |V|},$$

where $|V|$ denotes the vocabulary size of the model. This normalization bounds entropy values to $[0, 1]$ and prevents artificial inflation of uncertainty due to large vocabularies.

A.2 Centering and Power Transforms

Raw uncertainty signals often exhibit global bias or skewness that is unrelated to token-level decision difficulty. To mitigate this, optional centering operations are applied prior to scaling:

$$\psi(u_t) = u_t - \text{stat}(u),$$

where stat$(\cdot)$ denotes either the mean or median uncertainty computed over the answer token span.

In some configurations, a power or rectification transform $\phi(\cdot)$ is applied to emphasize high-uncertainty regions:

$$\phi(u) = \max(u, 0)^\gamma,$$

with $\gamma \in [1, 2]$. Unless otherwise stated, $\gamma = 1$ is used, corresponding to a linear mapping.

These operations are estimator-dependent and are applied consistently across all evaluated tokens within a given inference run.

A.3 Clipping and Stability Constraints

To prevent degenerate steering behavior, all token-level steering coefficients are bounded within fixed limits:

$$\alpha_t \in [\alpha_{\min}, \alpha_{\max}],$$

where $\alpha_{\min} = 0$ and $\alpha_{\max} = 2\alpha$ in all experiments unless explicitly varied.

Clipping ensures numerical stability and avoids excessive perturbations that could destabilize decoding, particularly under high-variance uncertainty signals.

A.4 Mean-Preserving Normalization

For multiplicative uncertainty-aware scaling, an additional normalization is applied to preserve the average steering strength across the answer span:

$$\alpha_t \leftarrow \alpha \cdot \frac{\alpha_t}{E_t[\alpha_t]}.$$

This constraint ensures that uncertainty-aware scaling redistributes steering strength across tokens rather than globally amplifying or suppressing the intervention. Additive and max-based formulations do not apply this normalization.

A.5 Signal-Specific Considerations

Different uncertainty estimators exhibit distinct statistical properties, requiring minor handling differences:

Predictive entropy. Entropy signals are smooth and bounded after vocabulary normalization. No additional smoothing is applied beyond optional centering.

Mahalanobis distance. Mahalanobis-based uncertainty is computed at the sequence level but applied uniformly to all tokens within an answer span. As a result, it does not induce token-wise variation but still allows sequence-level modulation of steering strength.

Recurrent Attention-Based Uncertainty (RAUQ). RAUQ aggregates uncertainty across selected attention heads and layers. In this work, only heads that exhibit consistent activation patterns across inference runs are retained. For LLaMA-2-7B-Chat, heads in layers 26 and 29 dominate the RAUQ signal and are therefore emphasized in the final uncertainty computation. RAUQ signals are normalized per layer prior to aggregation to prevent dominance by individual heads.