Master's Thesis Proposal

Fast, High-Recall Hallucination Flagging in Large Language Models via Weakly Supervised Anomaly Detection

1 Motivation

Large language models (LLMs) generate fluent text that can be incorrect, unverifiable, or unsupported by provided evidence—commonly called *hallucinations*. This remains a central blocker for real-world deployments such as RAG assistants, summarization pipelines, and decision-support systems, where a **recall-first safety gate** is often preferred: it is better to flag anything suspicious than to miss a hallucination.

Existing detection approaches force a trade-off between accuracy and cost. Multi-sample self-checking (e.g., SelfCheckGPT) achieves strong results but requires multiple LLM generations per input. Simple heuristics are fast but cover only narrow failure modes. Methods that rely on model internals (logits, hidden states) are inapplicable to closed-API models. This thesis targets the gap between these extremes: a detector that is **fast** (single-pass with lightweight post-processing), **high-recall** (with explicit target operating points), and **robust** across tasks, models, and RAG vs. non-RAG settings.

2 Problem Statement

Given a prompt x, optional evidence or context c, and an LLM answer y, the goal is to produce:

- a hallucination risk score $s(x, c, y)$,
- a binary flag decision $d \in \{0, 1\}$,

subject to three constraints:

- **Recall of hallucinations is maximized** under a latency budget.
- **Performance remains stable** under distribution shift (new domains, new models).
- Both **answer-level** and **sentence/claim-level** flags are supported.

3 Research Questions

The thesis is organized around three focused research questions:

RQ1 – Signal Effectiveness. Which cheap-to-compute signals (uncertainty proxies, semantic stability, retrieval support) most effectively separate hallucinated from grounded text, and how do they combine?

RQ2 – Weak Supervision vs. Unsupervised Detection. Does incorporating noisy or weak labels (from RAG evidence, self-consistency disagreement, or a small human-labeled set) improve recall stability and precision at fixed recall compared to a purely unsupervised anomaly baseline?

RQ3 – Recall-First Calibration. How can detection thresholds be set—and kept stable—to maintain a target recall (e.g., ≥90%) across domain and model shifts, and what is the resulting cost in false-positive rate?

4 Related Work

The proposal builds on three strands of prior work:

Sampling-based detection. SelfCheckGPT uses black-box inconsistency across multiple sampled generations to flag unsupported claims. It is effective but incurs significant latency due to repeated LLM calls. This thesis treats SelfCheckGPT as the primary quality baseline and aims to approach its recall at a fraction of its cost.

Uncertainty and entropy estimation. Token-level entropy and probability margins capture a subset of hallucinations, particularly fabricated entities and unsupported specifics. These signals form one component of the proposed feature set.

Benchmarks. TruthfulQA stress-tests model truthfulness. HaluEval provides hallucination recognition with human annotations. HalluLens disentangles hallucination from factuality and offers a clear taxonomy. RAG faithfulness evaluation surveys and metamorphic prompt-mutation approaches (MetaQA-style) complement these for the retrieval-augmented track.

5 Core Idea

The central thesis is that hallucinations can be detected as **anomalies in a feature space** constructed from cheap-to-compute signals extracted from (x, c, y). Rather than relying on any single signal or a heavy judge model, the approach combines a small set of complementary signals into a feature vector and applies anomaly detection on top.

The work proceeds in two stages:

Stage 1 – Unsupervised baseline. An unsupervised anomaly detector is trained on features from grounded (non-hallucinated) outputs, treating hallucinations as out-of-distribution.

Stage 2 – Weakly supervised refinement. Noisy positive labels (likely hallucinations) are introduced to sharpen the decision boundary. The key hypothesis is that even imperfect labels from cheap sources (RAG entailment failures, self-consistency disagreement) can substantially improve recall stability under distribution shift.

6 Methodology

6.1 Feature Design (Fast Signals)

Features are modular and budgeted so the full pipeline requires **zero or one extra LLM call**. They are grouped into four categories, each targeting a different hallucination mechanism:

A) Generation Uncertainty (grey-box, optional). Token-level entropy statistics (mean, max, slope) and top-1 vs. top-2 probability margins, with particular attention to entity and numeric spans. Available only when the serving API exposes logits.

B) Semantic Stability (black-box). Sensitivity of the answer to minor prompt perturbations (paraphrase, formatting changes), measured via embedding cosine similarity and lightweight NLI. This is a reduced-cost proxy for full self-consistency.

C) Retrieval Support (RAG track). Answer–context entailment score, entity and numeric novelty (claims not grounded in any retrieved passage), and approximate claim-to-evidence coverage.

D) Linguistic Cues. Overconfident phrasing without supporting evidence, ungrounded numeric specificity, and internal contradictions detected via NLI.

The output is an answer-level score, with optional sentence/claim-level scores for interpretability.

6.2 Detection Approach

Rather than exhaustively comparing many anomaly detection algorithms, the thesis focuses on a single, well-motivated pipeline with a clear unsupervised-to-weakly-supervised progression:

Unsupervised baseline. An Isolation Forest trained on feature vectors from grounded outputs serves as the primary unsupervised detector. Isolation Forest is chosen for its efficiency, interpretability, and strong empirical performance on tabular anomaly detection. A density-based alternative (e.g., a Gaussian Mixture Model) will be tested as a sanity check but is not the focus.

Weakly supervised refinement. The core contribution is to improve upon the unsupervised baseline by incorporating weak labels. The thesis will explore **PU (positive–unlabeled) learning** as the primary weak supervision framework, because it directly models the realistic setting where some hallucinations are cheaply flagged but the unlabeled set contains an unknown mix of grounded and hallucinated outputs. Noise-aware loss functions will handle label noise from the weak sources.

Recall-first calibration. Conformal prediction or quantile-based thresholding is applied on a held-out calibration set to guarantee a target recall (e.g., ≥90%) with a known false-positive rate, and its stability is evaluated across distribution shifts.

6.3 Weak Label Sources

Weak positive labels (likely hallucinations) are obtained from the following sources, ranked by cost:

- **RAG entailment failures:** claims that contradict or are unsupported by retrieved passages.
- **Self-consistency disagreement:** answers that change substantially across 2–3 sampled completions (low-budget variant of SelfCheckGPT).
- **Small human-labeled calibration set:** a targeted set of ~100–200 labeled examples for threshold tuning and validation.

LLM-as-a-judge soft scores may be used as an additional signal if the budget allows, but are not required.

7 Datasets and Evaluation

7.1 Datasets

Non-RAG hallucination: TruthfulQA, HaluEval, and HalluLens. **RAG hallucination:** standard RAG QA and summarization datasets evaluated with faithfulness protocols. **Stress testing:** MetaQA-inspired prompt mutations to generate controlled inconsistencies.

7.2 Baselines

SelfCheckGPT (quality ceiling), entropy/uncertainty-only baselines, simple RAG heuristics (entailment-only), and prompt-mutation baselines.

7.3 Metrics

The primary metric is **recall at fixed false-positive rates** (10%, 20%), reflecting the recall-first design goal. Secondary metrics include AUPRC, recall under latency budgets, and calibration quality (ECE, reliability curves).

7.4 Robustness Evaluation

Cross-domain transfer (train on one benchmark, evaluate on another), cross-model transfer (train on one LLM's outputs, evaluate on another's), and RAG-shift evaluation (change retriever or corpus). Feature ablations quantify the contribution of each signal group.

8 Expected Contributions

1. A **budgeted feature framework** for hallucination risk scoring that operates within a single extra LLM call (or zero for black-box features only).
2. An empirical comparison of **unsupervised vs. weakly supervised anomaly detection** for hallucination flagging, demonstrating when and why weak labels help.
3. A **recall-first calibration method** with measured stability under distribution shift.
4. An **open, reproducible evaluation pipeline** covering both RAG and non-RAG settings.

9 Engineering Plan

The implementation is Python-based, using scikit-learn for the Isolation Forest and PU learning components, and PyTorch where lightweight neural components are needed (e.g., NLI inference). The pipeline is modular: data loading, feature extraction, detector training, calibration, and evaluation are separate stages with standardized interfaces, enabling easy ablation and extension.

10 Risks and Mitigations

Definition ambiguity. "Hallucination" is used inconsistently across benchmarks. The thesis follows each benchmark's own taxonomy and reports results separately rather than conflating definitions.

Noisy weak labels. PU learning and noise-aware losses are specifically designed to handle label noise. The human-labeled calibration set provides a clean reference for threshold tuning.

High false-positive rate. A recall-first design necessarily incurs false positives. The thesis addresses this through tiered output (score + binary flag + explanation signals) so downstream consumers can choose their own operating point.

Closed-model limitations. The feature design maintains a fully black-box track (stability + linguistic cues) alongside the grey-box track (entropy features), so the approach works even without logit access.