Project Proposal: Self-Verification Chains for Hallucination-Free RAG

Team:
Hemanth Balla(UFID – 64798116)
Anisa Shaik(UFID – 49373585)
Reshma Koshy (UFID – 52903493)
Pranay Reddy Pullaiahgari (UFID – 62381134)

1. Problem Statement and Objectives

1.1 The Hallucination Problem

Retrieval-Augmented Generation (RAG) grounds LLM outputs in factual evidence, yet LLMs still hallucinate in 25-30% of open-domain QA responses even with retrieved context. **Core Problem:** Existing RAG systems retrieve documents and generate answers, but lack mechanisms to verify whether the generated answer is actually supported by the retrieved evidence, creating three critical gaps:

1. **No post-generation verification** - Systems cannot detect their own hallucinations
2. **No correction mechanism** - When hallucinations occur, there's no way to revise or reject answers
3. **No feedback loop** - Models don't improve from identifying and correcting errors

1.2 Our Solution

We propose a **self-verification RAG pipeline** treating hallucination suppression as a **measurable, modular process** with quantitative verification at each stage:

1. **Hybrid retrieval + reranking** - Ensure evidence quality (Recall@20 ≥ 0.95)
2. **Answer generation** - Maintain linguistic quality (F1 ≥ 0.58)
3. **Claim-level verification** - Check factual precision (≥ 0.90) using entailment models
4. **Adaptive revision** - Re-retrieve or regenerate when verification fails
5. **Iterative fine-tuning** - Use verified outputs (Factual Precision ≥ 0.85) as training data

1.3 Project Objectives

Primary Objectives:

- Build end-to-end self-verifying RAG pipeline with stage-specific quantitative metrics
- Achieve Hallucination Rate ≤ 0.10 (baseline ~0.25-0.30)
- Demonstrate Verified F1 (F1 × Factual Precision) improvement ≥ 20% vs. baseline
- Show iterative training reduces hallucination by ≥10% per iteration

Research Questions:

- RQ1: Can we achieve Recall@20 ≥ 0.95 with hybrid retrieval + reranking?
- RQ2: Does claim-level entailment verification achieve Factual Precision ≥ 0.90 and Recall ≥ 0.85?
- RQ3: What is the optimal entailment threshold τ that maximizes Verified F1?
- RQ4: Does iterative fine-tuning on verified outputs (score ≥ 0.85) reduce hallucinations over 3 iterations?

Success Metrics:

- **Retrieval:** Recall@20 ≥ 0.95, Coverage Index ≥ 0.90
- **Generation:** F1 ≥ 0.58, EM ≥ 0.43
- **Verification:** Factual Precision ≥ 0.90, Hallucination Rate ≤ 0.10
- **Composite:** Verified F1 ≥ 0.52 (20%+ improvement over baseline ~0.42)
- **Human Agreement:** ≥ 0.85 with automatic verifier labels

2. Proposed Methodology

2.1 System Architecture

```
User Question
        ↓
STAGE 1: Hybrid Retrieval
• Dense: FAISS + all-mpnet-base-v2
• Sparse: BM25 (k1=1.5, b=0.75)
• Fusion: 0.6 dense + 0.4 sparse
Metrics: Recall@20, MRR, NDCG@10
→ Top-50 passages
        ↓
STAGE 2: Cross-Encoder Reranking
• DeBERTa-v3-base on MS MARCO
• Score (Q, passage) relevance
Metrics: Coverage Index (≥ 0.90)
→ Top-10 passages
        ↓
STAGE 3: Answer Generation
• FLAN-T5-Large + QLoRA (r=16)
• Input: Q + Top-5 passages
Metrics: EM, F1, BLEU/ROUGE-L
→ Candidate answer
        ↓
STAGE 4: Claim-Level Verification
• Extract atomic claims (spaCy)
• NLI entailment (DeBERTa-MNLI + FEVER)
• Factuality = supported_claims / total
Metrics: Factual Precision, Recall,
         FEVER Score, FactCC
→ Factuality score [0-1]
        ↓
STAGE 5: Adaptive Decision & Revision
IF factuality ≥ 0.85: ACCEPT
IF 0.60-0.85: REVISE (re-retrieve)
IF < 0.60: REJECT (abstain)
Metrics: Abstention Rate, Revision Success
        ↓
STAGE 6: Iterative Fine-Tuning
• Collect verified outputs (≥0.85)
• Fine-tune on (Q, passages, answer)
Metrics: Per-iteration hallucination drop
→ Improved model
```

2.2 Component Details

Retrieval System:

- **Dense:** sentence-transformers/all-mpnet-base-v2 (768-dim) with FAISS IVF4096,PQ64 index (21M Wikipedia passages)
- **Sparse:** BM25 (k1=1.5, b=0.75) for keyword matching

- **Fusion:** Score = 0.6 × dense_score + 0.4 × sparse_score
- **Metrics:** Recall@5/10/20, MRR, NDCG@10
- **Target:** Recall@20 ≥ 0.95, MRR uptrend vs. dense-only baseline

Reranker:

- **Model:** microsoft/deberta-v3-base fine-tuned on MS MARCO passage ranking
- **Function:** Score (question, passage) pairs for answer containment
- **Metric:** Coverage Index = (answer tokens in retrieved docs) / (total answer tokens)
- **Target:** Coverage ≥ 0.90
- **Decision Rule:** If Recall@20 < 0.90 after reranking, re-train retriever

Generator:

- **Model:** google/flan-t5-large (780M params)
- **Fine-tuning:** QLoRA (4-bit NF4, rank=16, α=32, dropout=0.05)
- **Training:** SQuAD v2 + Natural Questions (130K examples, 3-5 epochs)
- **Prompt:** "Context: [passages]\nQuestion: [Q]\nBased strictly on context above, provide a factual answer. If insufficient evidence, respond 'Cannot answer based on context.'\nAnswer:"
- **Generation:** Beam search (k=5, length penalty=1.0)
- **Metrics:** EM, F1, BLEU-4, ROUGE-L, Abstention Rate

Verification Module:

- **Claim Extractor:** spaCy en_core_web_lg dependency parsing
  - Extract subject-verb-object triples
  - Filter factual statements (remove opinions, questions)
- **Entailment Checker:** microsoft/deberta-v3-large fine-tuned on MNLI + FEVER
  - For each claim: check if ANY passage entails it
  - Threshold: entailment_score > τ (default τ = 0.75)
  - Label: SUPPORTED / REFUTED / NO_EVIDENCE
- **Factuality Scoring:**
- Factual Precision = supported_claims / total_claims
- Factual Recall = supported_claims / gold_facts
- Hallucination Rate = 1 - Factual Precision
- FEVER Score = harmonic_mean(label_accuracy, evidence_recall)
- **Metrics:** Factual Precision ≥ 0.90, Factual Recall ≥ 0.85, Hallucination Rate ≤ 0.10

Revision Strategies:

1. **Answer-aware re-retrieval:** Use generated_answer + question as new query, retrieve additional top-5 passages for unsupported claims
2. **Constrained generation:** Regenerate with T=0.3, prompt: "Cite evidence for each claim"
3. **Claim-by-claim:** Regenerate only unsupported portions, merge with verified parts

Iterative Training Loop:

- Collect answers with Factual Precision ≥ 0.85 as verified training data

- Fine-tune generator on (question, top-5 passages, verified_answer) triples
- Run for 3 iterations, expect ≥10% hallucination reduction per iteration
- Monitor data diversity: lexical variety, syntactic complexity

2.3 Datasets

Training & Evaluation:

- **SQuAD v2.0:** 150K questions (includes unanswerable for abstention testing)
- **Natural Questions:** 307K real Google queries
- **HotpotQA:** 113K multi-hop questions (out-of-distribution stress test)
- **FEVER:** 185K fact verification claims (train verification module)

Corpus: Wikipedia 2018 dump (21M passages, pre-indexed with FAISS)

Split: Train 70% / Validation 15% / Test-ID 10% / Test-OOD 5%

2.4 Implementation

Frameworks: PyTorch 2.1+, Transformers 4.35+, PEFT 0.7+, FAISS, Sentence-Transformers, DeepSpeed, bitsandbytes

Compute: HiperGator (NVIDIA A100/V100), ~60-80 GPU hours total

Hyperparameters:

- Retrieval: Hybrid fusion 0.6/0.4, top-50 → rerank to top-10 → select top-5
- Generation: LR=1e-5, batch=4 (grad_accum=8), LoRA r=16, epochs=3-5
- Verification: τ=0.75 (tunable), accept≥0.85, revise 0.60-0.84, reject<0.60

3. Planned Experiments & Evaluation

3.1 Quantitative Evaluation Framework

We treat hallucination suppression as a **measurable, modular process** with stage-specific metrics:

Stage 1: Retrieval Performance

Purpose: Ensure evidence quality before generation

Metric	**Definition**	**Target**
Recall@20	Fraction of gold-supporting passages in top-20	≥ 0.95
MRR	Mean Reciprocal Rank of first relevant doc	Uptrend vs baseline
NDCG@10	Normalized Discounted Cumulative Gain (ranking quality)	> 0.80
Coverage Index	% of answer tokens linked to retrieved docs	≥ 0.90

Decision Rule: If Recall@20 < 0.90 → Re-train retriever or strengthen reranking

Stage 2: Generation Performance

Purpose: Confirm linguistic and answer quality before factual check

Metric	**Description**	**Notes**

Exact Match (EM)	Binary exact answer match	Standard for SQuAD/NQ
F1	Token overlap between generated and reference	Baseline quality measure
BLEU-4 / ROUGE-L	Text similarity for multi-sentence answers	Use for HotpotQA
Abstention Rate	% of "insufficient evidence" responses	Should ↑ as verification strengthens

Analysis: Compare EM/F1 before and after verification. EM should stay similar while hallucination rate drops.

Stage 3: Factual Verification Performance

Purpose: Measure hallucination suppression effectiveness

Metric	**Computation**	**Goal**
Factual Precision	True-factual claims / all claims	≥ 0.90
Factual Recall	True-factual claims / all gold facts	≥ 0.85
Hallucination Rate	1 - Factual Precision	≤ 0.10
FEVER Score	Harmonic mean(label accuracy × evidence recall)	Benchmark vs FEVER baselines
FactCC Score	Correlation with human factual judgments	Use pretrained FactCC if available

Verification Pipeline: Answer → atomic claims → NLI entailment vs evidence → aggregate (entailment > τ) as factual

Stage 4: End-to-End Composite Metrics

Verified F1 = F1 × Factual Precision

This composite shows factuality AND quality together:

- **Baseline RAG (no verifier):** F1 = 0.60, Factual Precision = 0.70 → Verified F1 = 0.42
- **Our Verified RAG:** F1 = 0.58, Factual Precision = 0.92 → Verified F1 = 0.53 (+26%)

Target: Verified F1 improvement ≥ 20% over baseline

Stage 5: Human Validation (100 samples)

For each answer, annotator labels:

- **SUPPORTED:** All claims backed by retrieved evidence
- **CONTRADICTED:** Contains claims contradicting evidence
- **NO EVIDENCE:** Contains unsupported claims (hallucination)

Metric: Human-Verifier Agreement = % of matching labels

Target: ≥ 0.85 agreement → automatic verifier is reliable

Stage 6: Logging for Traceability

Every experiment logs:

- Dataset: name, split, commit hash (e.g., squad-v2-dev-a3f7b2)
- Retriever: model version, FAISS index ID, build timestamp
- Generator: checkpoint path, training iteration number

- Verifier: model name, threshold τ
- All metrics: EM, F1, Recall@k, MRR, NDCG, Coverage, Factual Precision/Recall, FEVER, FactCC, Verified F1, Abstention Rate

Storage: W&B run with tagged artifacts + local JSON backup
Purpose: Instructor can reproduce any result deterministically

3.2 Experiment Design
Experiment 1: Baseline RAG Performance
Goal: Establish baseline without verification
Setup: Standard retrieve (FAISS, top-5) → generate (FLAN-T5), no verification/reranking/revision
Metrics:

- Generation: EM, F1, BLEU/ROUGE-L
- Factuality: Manual annotation on 100 samples → Factual Precision, Hallucination Rate
- Composite: Verified F1

Expected: EM ~40-45%, F1 ~55-60%, Factual Precision ~0.70, Verified F1 ~0.42

Experiment 2: Retrieval Strategy Comparison
Goal: Quantify impact of retrieval method and reranking on evidence quality
Configurations:

- A: Dense only (FAISS, all-mpnet-base-v2)
- B: Sparse only (BM25)
- C: Hybrid (0.6 dense + 0.4 sparse)
- D: Hybrid + Cross-encoder reranking (DeBERTa-v3-base)

Metrics:

- Retrieval: Recall@5/10/20, MRR, NDCG@10, Coverage Index
- Downstream: QA F1, Factual Precision (measure on 100 samples per config)

Hypothesis: Config D achieves Recall@20 ≥ 0.95, Coverage ≥ 0.90, improves downstream Factual Precision by 8-12%
Decision: If Recall@20 < 0.90 in Config D, iterate on reranker fine-tuning

Experiment 3: Verification Threshold Tuning
Goal: Find optimal entailment threshold τ that maximizes Verified F1
Variables: τ = {0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90}
For each τ, measure:

- Factual Precision (higher τ → stricter → higher precision)
- Answer Recall (higher τ → more rejections → lower recall)
- Verified F1 = F1 × Factual Precision
- Abstention Rate (should increase with τ)

Visualization: Plot Factual Precision vs. Answer Recall curve

Analysis: Select τ that maximizes Verified F1 while maintaining Factual Precision ≥ 0.90
Expected Optimal: $\tau = 0.75\text{-}0.80$

Experiment 4: Revision Strategy Evaluation
Goal: Compare revision approaches when verification triggers ($0.60 \leq$ factuality < 0.85)
Strategies:

- A: No revision (accept/reject only)
- B: Answer-aware re-retrieval (use answer + Q as new query)
- C: Constrained generation (strict prompt, T=0.3, cite evidence)
- D: Claim-by-claim regeneration (fix only unsupported claims)

Metrics:

- Revision Success Rate: % of revised answers with improved factuality
- Final Verified F1 after revision
- Compute cost: # extra retrieval/generation calls

Hypothesis: Strategy B (re-retrieval) most effective, 10-15% Verified F1 improvement with 1.5× compute cost

Experiment 5: Generation Strategy Comparison
Goal: Evaluate different decoding methods for hallucination reduction
Configurations:

- A: Greedy decoding (baseline)
- B: Beam search (k=5, length penalty=1.0)
- C: Self-consistency (generate 5 samples at T=0.7, verify each, majority vote)

Metrics:

- Hallucination Rate per strategy
- EM, F1 (answer quality)
- Verified F1
- Computational cost (C requires 5× generation)

Hypothesis: Self-consistency + verification reduces Hallucination Rate by 15-20% but costs 5× compute

Experiment 6: Iterative Fine-Tuning Loop
Goal: Test if verified outputs improve model over iterations
Process:

1. **Iteration 0:** Train FLAN-T5 on SQuAD v2 (baseline)
2. **Iteration 1:** Generate 10K answers on train set, verify, collect 5K with Factual Precision ≥ 0.85, fine-tune
3. **Iteration 2:** Repeat with improved model
4. **Iteration 3:** Final iteration

Metrics per iteration:

- Hallucination Rate (should decrease ≥10% per iteration)
- QA F1, EM (should stay stable or improve slightly)
- Verified F1 (should increase each iteration)
- Training data quality: avg Factual Precision of collected examples

Success Criteria:

- Iteration 1: Hallucination Rate drops to ≤0.18 (from ~0.25)
- Iteration 2: ≤0.12
- Iteration 3: ≤0.10

Monitor: Data diversity (lexical variety using type-token ratio, syntactic complexity)

Experiment 7: Component Ablation Study

Goal: Quantify contribution of each pipeline component

Ablations:

- Remove reranking: Direct FAISS → generation
- Remove verification: Standard RAG (no factuality check)
- Remove revision: Verify but don't retry when fails
- Simpler verifier: Lexical overlap instead of DeBERTa-NLI

Metrics: For each ablation, measure drop in:

- Verified F1
- Factual Precision
- Hallucination Rate increase

Analysis: Identify most critical components (expect verification to have largest impact)

Experiment 8: Stress Testing & Pareto Analysis

Goal: Validate robustness and visualize quality-factuality tradeoff

Test 1: Threshold Sweep (from Exp 3)

- Plot Pareto frontier: X-axis = EM, Y-axis = (1 - Hallucination Rate)
- Show curves for: Baseline RAG, Verified RAG at different τ
- Demonstrate: Verified RAG dominates (higher on both axes at optimal τ)

Test 2: Retrieval Degradation

- Artificially degrade: Set Recall@20 = {0.95, 0.85, 0.75, 0.65}
- Measure: Downstream Factual Precision drop
- Validate: Poor retrieval → poor factuality (as expected)

Test 3: Verifier Off

- Turn off verification entirely
- Measure: Hallucination Rate increase (should be +15-25% vs. verified)
- Confirms: Verifier is essential

Visualization: 3-panel plot:

- Panel A: Threshold vs. Factual Precision & Recall
- Panel B: Recall@20 vs. Downstream Factual Precision

- Panel C: Pareto frontier (EM vs. Factuality)

3.3 Statistical Rigor

For all experiments:

- Run 3 times with different random seeds (42, 123, 456)
- Report: Mean ± standard deviation
- Significance testing: Paired t-test ($p < 0.05$) for pairwise comparisons
- Bootstrap resampling (1000 iterations) for confidence intervals on human evaluation

4. Team Member Contributions

4.1 Individual Responsibilities

Hemanth Balla - Retrieval & Iterative Training Lead

Tasks:

- Implement dual retrieval (FAISS + BM25 hybrid), optimize fusion weights
- Fine-tune cross-encoder reranker on MS MARCO
- Measure Recall@k, MRR, NDCG@10, Coverage Index for all retrieval configs
- Build iterative fine-tuning loop: collect verified data (≥0.85), retrain generator
- Run Experiments 2 (retrieval comparison) and 6 (iterative training)
- Statistical analysis: paired t-tests, bootstrap confidence intervals
- Log all retriever versions, index IDs, checkpoints for reproducibility

Deliverables: Retrieval pipeline, reranker, iteration loop, Exp 2&6 results

Anisa Shaik - Generation & Verification Lead

Tasks:

- Fine-tune FLAN-T5-Large with QLoRA (r=16, 4-bit quantization)
- Train DeBERTa-v3-large verifier on MNLI + FEVER datasets
- Implement claim extraction (spaCy), entailment checking, factuality scoring
- Build revision strategies (re-retrieval, constrained generation, claim-by-claim)
- Integrate all components into end-to-end pipeline
- Run Experiments 4 (revision) and 5 (generation strategies)
- Compute FEVER Score, FactCC Score for all verified outputs
- Log generator checkpoints, verifier thresholds, revision parameters

Deliverables: Generator model, verifier, revision module, end-to-end pipeline, Exp 4&5 results

Reshma Koshy - Data & Evaluation Lead

Tasks:

- Download and preprocess SQuAD v2, NQ, HotpotQA, FEVER datasets
- Build FAISS index on Wikipedia (21M passages), version and timestamp
- Implement claim extraction module (spaCy dependency parsing)
- Design human evaluation rubric (SUPPORTED/CONTRADICTED/NO_EVIDENCE)

- Coordinate 100-sample human annotation study, compute inter-annotator agreement
- Compute Human-Verifier Agreement (target ≥0.85)
- Manage verified data collection for iterative training (track diversity metrics)
- Maintain experiment logs: dataset commits, index IDs, all metric CSVs

Deliverables: Clean datasets, FAISS index, claim extractor, human eval data (100 samples), verified data for training

Pranay - Experiments & Analysis Lead
Tasks:
- Set up evaluation infrastructure: metric computation, logging, W&B integration
- Implement self-consistency generation (5 samples, T=0.7, voting)
- Run Experiments 1 (baseline), 3 (threshold tuning), 7 (ablations), 8 (stress testing)
- Compute all metrics: EM, F1, BLEU, ROUGE-L, Factual Precision/Recall, Hallucination Rate, FEVER Score, FactCC, Verified F1
- Generate visualizations:
    - Threshold sweep (Factual Precision vs. Recall)
    - Pareto frontier (EM vs. Factuality)
    - Iteration curves (Hallucination Rate over 3 iterations)
    - Ablation bar charts (component contribution)
- Error analysis: categorize hallucination types (fabrication, extrapolation, misattribution)
- Maintain deterministic logging: commit hashes, model versions, all configs

Deliverables: Evaluation scripts, all metric CSVs, visualizations (4+ figures), ablation results, stress test analysis

5. Expected Contributions and Impact
5.1 Novel Contributions
1. **Modular quantitative framework:** First work to decompose RAG hallucination mitigation into stage-specific measurable metrics (Retrieval: Recall@20/Coverage, Generation: EM/F1, Verification: Factual Precision/FEVER Score)
2. **Verified F1 composite metric:** New evaluation metric (F1 × Factual Precision) that jointly captures answer quality and factuality, enabling apples-to-apples comparison across RAG systems
3. **Iterative self-improvement:** Demonstrate that verified outputs (Factual Precision ≥ 0.85) create high-quality training data, enabling 3-iteration self-improvement with ≥10% hallucination reduction per cycle
4. **Threshold optimization framework:** Systematic analysis of entailment threshold τ vs. Pareto frontier (EM vs. Factuality), providing practitioners with evidence-based threshold selection

5.2 Comparison to Related Work

- **Standard RAG (Lewis et al., 2020):** No verification → **Our addition:** Claim-level verification + revision + iterative training
- **SelfCheckGPT (Manakul et al., 2023):** Self-consistency without retrieval → **Our difference:** RAG setting with evidence-based verification and quantitative metrics
- **RARR (Gao et al., 2023):** Retrieve, revise, read → **Our difference:** Claim-level verification, iterative fine-tuning loop, stage-specific quantitative metrics

8. References

1. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*.
2. Manakul, P., et al. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *arXiv:2303.08896*.
3. Gao, Y., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997*.
4. Thorne, J., et al. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. *NAACL*.
5. Rajpurkar, P., et al. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *ACL*.
6. Kwiatkowski, T., et al. (2019). Natural Questions: A Benchmark for Question Answering Research. *TACL*.
7. Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*.
8. Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*.
9. Nogueira, R., et al. (2020). Document Ranking with a Pretrained Sequence-to-Sequence Model. *EMNLP*.
10. Kryscinski, W., et al. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. *EMNLP* (FactCC).