Technical Document: Improving Safety and Reliability of Conversational AI

Executive Summary

This document analyzes four critical issues in conversational AI and proposes solutions for the two highest priorities: **hallucination** and **bias**.

---

1. Problem Analysis

1.1 Inconsistent Responses

Causes: Lack of explicit memory, position encoding decay, training objective mismatch, context truncation.
Measurement: NLI contradiction detection, consistency probes, self-verification.

1.2 Hallucination ⭐ Priority 1

Causes: Training data noise, maximum likelihood objective, lack of grounding, knowledge cutoff.
Measurement: TruthfulQA, FEVER, citation verification, calibration error (ECE).

1.3 Bias ⭐ Priority 2

Causes: Training data bias, representation imbalance, annotation bias, stereotype amplification.
Measurement: BBQ, WinoBias, StereoSet, counterfactual analysis.

1.4 Prompt Sensitivity

Causes: Distributional shift, attention sensitivity, limited instruction following.
Measurement: Paraphrase consistency, perturbation analysis.

Prioritization Rationale

Issue	Severity	Priority

Issue	Severity	Priority
Hallucination	High (trust erosion)	**1**
Bias	High (harm to groups)	**2**
Inconsistency	Medium	3
Prompt Sensitivity	Medium	4

---

2. Proposed Solutions

2.1 Hallucination Mitigation: RAG + Uncertainty Estimation

Architecture:

```
Query → Retriever → Top-K Docs → [Query+Docs] → LLM → Response + Citations + Confidence → Verifier
```

Components:

1. Dense retrieval from curated knowledge base
2. Attribution mechanism with inline citations
3. Uncertainty head for confidence estimation
4. Post-generation verification

Resources: 64 A100 GPUs, 6 months, ~100GB knowledge base

Success Metrics:

- 50% reduction in hallucination rate
- <10% calibration error
- > 85% citation accuracy

2.2 Bias Mitigation: Multi-Stage Pipeline

Stages:

1. **Data Curation:** Audit, filter, augment with counter-stereotypical examples
2. **Training:** Contrastive debiasing, adversarial training, balanced sampling
3. **Inference:** Bias classifier, prompt augmentation, output editing

Resources: 32 A100 GPUs, 6 months

Success Metrics:

- 40% reduction in StereoSet bias
- <5% demographic parity difference

---

3. Experimental Design: Hallucination Mitigation

Hypothesis

RAG + uncertainty estimation reduces hallucination by ≥40% while maintaining quality within 5%.

Setup

- **Control:** Baseline model
- **Treatment:** RAG-only, Uncertainty-only, Full system
- **Datasets:** TruthfulQA, FEVER, Natural Questions (~1000 examples)

Analysis

- Two-proportion z-test for hallucination rates
- Bootstrap 95% CIs
- Bonferroni correction

Interpretation

Outcome	Action
>40% reduction	Deploy with monitoring
20-40% reduction	Iterate on retrieval
<20% reduction	Investigate failures

---

4. Broader Implications

Trade-offs

Intervention	Safety Benefit	Cost
Retrieval	Grounds claims	+100-200ms latency

Intervention	Safety Benefit	Cost
Uncertainty	Flags unreliable outputs	May over-refuse

User Communication

- Visual confidence indicators
- Source citations
- Clear changelog and limitations documentation

References

1. Lewis et al. (2020) "Retrieval-Augmented Generation" NeurIPS
2. Lin et al. (2022) "TruthfulQA" ACL
3. Parrish et al. (2022) "BBQ" ACL Findings