ICLR 2026 · Workshop on Logical Reasoning of LLMs

From Facts to Conclusions: Integrating Deductive Reasoning in Retrieval-Augmented LLMs

Shubham Mishra · Shiv Tiwari · Samyek Jain · Gorang Mehrishi · Dhruv Kumar · Pratik Narang · Harsh Sharma

BITS Pilani

Carnegie Mellon University

KEY RESULTS

0.069 → 0.883

Answer Correctness

+1,179%

0.074 → 0.722

Behavioral Adherence

+876%

F1 = 1.0

F1-GR (Grounded Refusal)

Perfect

github.com/ShubhamX90/rag_reason

arXiv: 2512.16795

Agenda

01

Paper: Problem & Framework

RAG fails under conflict · 3-stage reasoning pipeline

02

Paper: Dataset & Results

539-query benchmark · CATS metric · SFT gains

03

New: Annotation Pipeline v3

Multi-strategy annotation · Multi-LLM committee · Batch mode

★ NOT IN PAPER

04

New: CATS v2.0

Multi-judge voting · Enhanced metrics · Cost optimization

05

New: SFT Pipeline v2

NEFTune · Mixed-mode training · Improved LR schedule

06

Roadmap & Future Work

Next steps across all three components

01 Problem Statement

Why RAG Fails in the Real World

Conflicting Sources

Different documents assert incompatible facts for the same query

Outdated Information

Retrieved docs contain stale data that contradict newer evidence

Incomplete Evidence

No single doc fully answers the query — models fill gaps by hallucinating

No Justified Refusal

Models answer even when they should abstain — dangerous in practice

The Critical Gap

No existing framework simultaneously addresses conflict detection, structured reasoning, grounded citation, and justified refusal.

What Prior Work Misses

Cattan et al. 2025

Conflict taxonomy — but no training to resolve conflicts

Chain-of-Note 2024

Per-doc grounding — but ignores cross-doc disagreement

Trust-Score 2025

RAG fidelity metric — but no behavioral alignment

01 3-Stage Reasoning Pipeline

Stage 1

Micro Reasoning

Each doc labeled: supports / partially supports / irrelevant
Key fact extracted · Source quality scored · Quote anchored

Stage 2

Conflict Analysis

CB

Aggregates verdicts → infers one of 5 conflict types
Generates concise conflict rationale (≤60 words)

Stage 3

Grounded Synthesis

Citation-linked answer or justified refusal
<think> XML trace · Conflict-aware behavior

5 Conflict Types Handled:

- No Conflict
- Complementary Info
- Conflicting Opinions
- Outdated Info
- Misinformation

`<think>` All 3 stages serialized as XML reasoning traces · Transparent · Interpretable · Citation-linked · Enables conflict-sensitive SFT supervision `</think>`

02 Dataset, CATS Metric & Results

539	5	80/10/10	54
Queries	Conflict Types	Train/Val/Test	Test Queries

CATS — Conflict-Aware Trust Score

F1-GR — **Grounded Refusal**
Answerable vs. abstain

AC — **Answer Correctness**
From docs only, no hallucination

GC — **Grounded Citation**
NLI entailment verification

BA — **Behavioral Adherence**
LLM-as-Judge (GPT-4o)

End-to-End Results (SFT vs Baseline)

Model	F1-GR	Ans. Corr	Gnd. Cit	Behav.
Mistral Base	0.870	0.604	0.515	0.630
Mistral SFT	**1.000** ✓	**0.930**	**0.678**	**0.741**
Qwen Base	0.167	0.069	0.111	0.074
Qwen SFT	**1.000** ✓	**0.883**	**0.648**	**0.722**

Qwen SFT: **Answer Correctness 0.069 → 0.883 (+1,179%)** | **Behavioral Adherence 0.074 → 0.722 (+876%)** | **F1-GR = 1.0 ✓**

Beyond the Paper

Three major engineering components built after submission

★ NEW

Annotation Pipeline v3

Multi-strategy · Multi-LLM committee · Batch mode

★ NEW

CATS v2.0

Multi-judge voting · Enhanced metrics · Cost tracking

★ NEW

SFT Pipeline v2

NEFTune · Mixed-mode training · Better LR schedule

03 ★ NEW Dataset Annotation Pipeline v3

Strategy A: 3-Stage

Best for: Maximum traceability · Stage-by-stage debugging

1. Stage 1: N calls per query (one per doc)
   Per-doc verdict + evidence extraction
2. Stage 2: 1 call per query
   Conflict reasoning + answerability
3. Stage 3: 1 call per query
   Grounded response + <think> trace

Strategy B: Monolithic

Best for: Fewer round trips · Faster end-to-end

1. 1 call per query
2. All stages combined: per-doc verdicts + conflict reason + response
3. Same output schema as 3-Stage — fully interchangeable downstream

LLM Approaches — Works with BOTH strategies

Single-LLM

One model runs chosen strategy
Claude Sonnet 4.6, GPT-4o, OpenRouter/Qwen

Multi-LLM Committee

Voting across multiple models via OpenRouter
Weighted majority · Confidence scoring

Execution Modes

Async (fast, concurrent)
Batch (Anthropic/OpenAI discounted API)

04 ★ NEW CATS v2.0 — Conflict-Aware Trust Score

What's New in v2.0

Multi-LLM Judge Committee

Reduces single-judge bias · Improves reliability through consensus · Configurable voting strategies

Enhanced Factual Grounding

Cross-document verification · Claim-level breakdown · Supporting doc tracking

Improved Single-Truth Recall

Semantic matching beyond string matching · Paraphrase detection · Partial credit

Async Architecture

10× faster with parallel judge execution · Non-blocking API calls

Cost Tracking

Per-model cost breakdown in evaluation report · Token-level accuracy

Typical cost: $0.005–0.01/sample (default) · $0.005–0.01 (conservative, free tier) · $0.003–0.005 (single judge)

Multi-LLM Judge Committee

Claude Haiku 3.5 Anthropic	Fast, high-quality	$$$
DeepSeek R1 OpenRouter	Strong reasoning	$$$
Qwen 3 8B OpenRouter	Balanced performance	$$
Mistral Nemo OpenRouter	Zero cost, good quality	FREE

Weighted Majority Voting Strategy

Each judge evaluates independently (async) · Votes weighted by priority + confidence · Threshold: 0.6 · Rationale from top-weighted winner

05 ★ NEW SFT & Inference Pipeline v2 — Key Improvements

Aspect	v1 (Paper)	v2 (New)
Prompting	Stage-wise (3 calls) + Monolithic + Ablations + Simple RAG	**Single unified prompt (monolithic only)**
Oracle Levels	oracle1 (conflict type), oracle2 (per-doc), oracle3 (both)	**oracle (conflict type only) — cleaner**
Training Prompts	Separate training/inference prompts	**Same prompts for training + SFT inference + baseline**
Fine-Tuning	Basic QLoRA	**QLoRA + NEFTune + better LR + mixed-mode + class weighting**

N

E

NEFTune

Noisy Embedding Fine-Tuning

Adds uniform noise to token embeddings during training. Empirically improves SFT performance by 2-8%.

M

M

Mixed-Mode Training

E2E + Oracle data combined

Train on both E2E and Oracle message files simultaneously. Doubles effective dataset size, improves generalization.

LR

Better LR Schedule

Cosine decay + warmup

6% linear warmup → cosine decay. Weight decay (0.01) for regularization. Gradient norm clipping (max 1.0).

CB

Class Balancing

Improved weighting strategy

Weighted random sampler with 1/√count per class. Conflict label tokens up-weighted by conflict_weight=3.0.

06 Roadmap & Future Work

Research

- Evaluate on larger conflict benchmarks beyond CONFLICTS dataset
- Strengthen conflict-type supervision for Conflicting Opinions
- Integrate retriever-aware signals for end-to-end robustness
- Extend to multi-turn and real-world decision-support settings

Annotation Pipeline v3

- Add support for OpenAI batch mode in multi-LLM committee path
- Resume support for individual stage failures mid-run
- Cross-strategy comparison dashboard (3-Stage vs Monolithic)
- Automated quality gates before stage transitions

CATS v2.0

- Add claim-level grounding breakdown to main report
- Support additional OpenRouter free-tier models in committee
- Cross-run comparison reports for ablation studies
- Web UI for interactive evaluation result exploration

SFT Pipeline v2

- Support LLaMA 3.1 / Mistral v0.3 as base models
- DPO / PPO reward shaping for conflict-aware behavior
- Multi-GPU training with DeepSpeed ZeRO
- Automatic best-checkpoint selection via CATS v2.0 score

Thank You

From Facts to Conclusions: Integrating Deductive Reasoning in Retrieval-Augmented LLMs

ICLR 2026 · Workshop on Logical Reasoning of Large Language Models

Paper arXiv: 2512.16795

Code github.com/ShubhamX90/rag_reason

Contact f20220763@pilani.bits-pilani.ac.in

Contributions

- Conflict-aware RAG framework (paper)
- 539-query reasoning dataset + CATS metric
- SFT gains: +1,179% answer correctness
- Annotation Pipeline v3 (new)
- CATS v2.0 with multi-judge committee
- SFT Pipeline v2 with NEFTune