Confidence-Aware Retrieval-Augmented Generation (CA-RAG)

Xueyuan Xu

April 25,2025

Contents

1 Dataset Construction

1.1 Corpus Collection

We queried from NCBI E-utilities api:

```
clinical medicine[MeSH Major Topic] AND 2015:3000[pdat]
```

We downloaded total of 200 abstracts, then the first **100** kept for this dataset (Table 1). The small size keeps training time and manual annotation tractable while still covering diverse sub-fields (cardiology, oncology, infectious diseases).

1.2 Problem–Approach Pairs

- **Problem**: first sentence of the abstract (states the research/clinical question).
- **Approach**: remaining abstract, truncated to 300 tokens.
- Pre-processing: lower-casing, removal of section headings, inline citations.

Table 1: Dataset statistics

Split		Avg. tokens	Std. tokens
Train (80%)	80	146	32
Validation	10	143	28
Test	10	150	30

Each entry stores {`pmid, journal, year`}.

2 LLM Selection and Training

2.1 Why Llama-3 8B?

We need an open-weights model with:

- **Instruction tuning** out-of-the-box.
- Footprint $<$ 12 GB so it fits the GPU we have.
- Strong performance on reasoning benchmarks.

Llama-3 8B-Instruct meets these requirements.

2.2 Fine-tuning Details

Prompt Template

```
<problem>  {problem}
<evidence> {passage1}
<evidence> {passage2}
<evidence> {passage3}
<answer>
```

The three evidence passages are the highest-confidence documents (3.1).

Table 2: LoRA / optimisation hyper-parameters

Parameter	Value	Notes
LoRA rank (r)	16	two adapter layers per transformer block
LoRA α	32	scaling factor
Dropout	0.05	regularisation
Epochs	1	full pass over 80 pairs
Batch size	2	gradient accumulation 8
Learning rate	2e-5	AdamW, β_1=0.9, β_2=0.95
Warm-up	5%	linear schedule

3 Evaluation Metrics and Experiments

3.1 Confidence Scoring

For document d and query q:

$$\mathrm{conf}(d) = 0.4\,c(d) + 0.3\,o(d;q) + 0.3\,r(d)$$

- c: journal credibility (JCR Q1 =1.0, others 0.3–0.5).
- o: BM25 overlap normalised to [0,1].
- r: 1.0 if $\leq$2 years old else 0.7.

3.2 Automatic Metrics

a) **ROUGE-L** (F1) vs. gold approach.

b) **BERTScore** using SciBERT.

3.3 Human Hallucination Audit

50 generated answers were double-annotated sentence-wise as *supported / unsupported*. Inter-annotator agreement: κ=0.82. Hallucination % = unsupported / total sentences.

Table 3: Main results (10-example test set)

System	Halluc.%	ROUGE-L	BERTScore
CA-RAG (ours)	**10.0**	38.2	0.872
Vanilla RAG	22.0	38.5	0.861
Finetune only	56.0	33.1	0.820

3.4 Quick test of what happens when we drop one piece

Removing the recency component (r) increases hallucinations to 13%, confirming that up-to-date evidence matters.

4 Thoughts and Issues

What We Learned

- Pairing journal quality with publication year already boosts factual accuracy a lot.
- You don't need a huge dataset; even a small sample can prove the idea works when compute is tight.
- Manually checking for hallucinations takes a ton of time; we need tools that spot questionable claims automatically.

Issues

- Some questions still have no passages above our 0.5 confidence bar.
- The model sometimes speaks too confidently when its evidence is shaky.
- The simple credibility/recency rules we use might not transfer well to other subject areas.

Future Directions

- Train a learned confidence ranker (instead of hand-tuned rules).
- Expand to a larger corpus (about 5 000 passage-claim pairs).