AIC CTU@FEVER 8: On-premise fact checking through long context RAG

Herbert Ullrich
AI Center @ CTU FEE
Charles Square 13
Prague, Czech Republic
ullriher@fel.cvut.cz

Jan Drchal
AI Center @ CTU FEE
Charles Square 13
Prague, Czech Republic
drchajan@fel.cvut.cz

Abstract

In this paper, we present our fact-checking pipeline which has scored first in FEVER 8 shared task. Our fact-checking system is a simple two-step RAG pipeline based on our last year's submission. We show how the pipeline can be redeployed on-premise, achieving state-of-the-art fact-checking performance (in sense of Ev^{2}R test-score), even under the constraint of a single Nvidia A10 GPU, 23GB of graphical memory and 60s running time per claim.

1 Introduction

In 2024, Automated Verification of Textual Claims (AVeriTeC) shared task (Schlichtkrull et al., 2024a) showed that the fact checking of real-world claims like those from Politifact, AfricaCheck, etc., can be automated to a significant extent, with pipelines accessing Large Language Models (LLMs) to produce the evidence and veracity verdicts for previously unseen claims instead of a human. Almost each competitive AVeriTeC shared-task system, however, relied on a proprietary LLM like GPT-4o (Rothermel et al., 2024; Ullrich et al., 2024) or an open-weights model with high tens of billions of parameters (Yoon et al., 2024). This raised a concern – can the fact-checking process be automated in a way accessible to masses, or is its quality conditioned by the business-owned blackbox models or access to prohibitive computational resources?

In this year's FEVER 8 shared task, the challenge is to match the quality of AVeriTeC systems with ones that only use open-weights models, constrained time of 60 seconds per claim on average, and a fixed compute of a single 23GB A10 GPU.

Our AIC CTU system (Figure 1), adapted for FEVER 8 from our last year submission, tops its test-leaderboard (Table 1) with a simple Retrieval-augmented Generation (RAG) scheme, using a locally hosted (Ollama) instance of Qwen3 LLM with 14B parameters, leveraging the sheer context length modern-day LLMs can process.

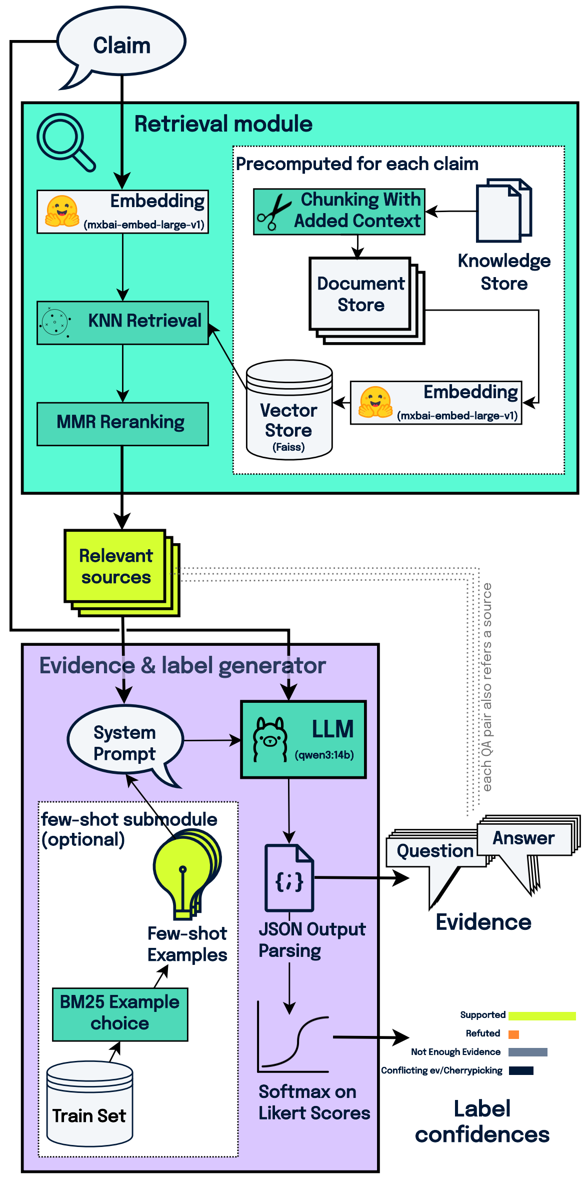

Figure 1: Our refreshed fact-checking pipeline used in CTU AIC FEVER 8 submission, adapted from Ullrich et al. 2024.

This paper introduces our system, discusses its design choices and how do they account on the score. We suggest our system as the new strong baseline – simple at core, competitive results – providing the code and reproduction advice.

2 System description

Our system is a straightforward adaptation of the AIC CTU Averitec system designed one year prior, published in Ullrich et al. 2024. The cited paper describes the system in detail, with ablation studies and justifications of each step. Our pipeline, depicted in Figure 1, consists of precomputation, retrieval, and generation modules:

i. Precomputation module

1. The provided AVeriTeC **knowledge store** (Schlichtkrull et al., 2024b) is split into chunks of specified maximum length, each marked with metadata of its URL and the full texts of the chunk before and after.
2. The chunks are then embedded into their vector representations, using only the chunk texts and no metadata.
3. Out of all chunk embeddings, a **vector store** is produced for each claim to be stored as a vector database.

ii. Retrieval module

1. The **Claim** is embedded into its vector representation using the same model used in i.2.
2. k nearest neighbours are then retrieved from the vector store, along with their **chunk embeddings**
3. The chunk embeddings are then reranked using the Maximal Marginal Relevance (MMR) method (Carbonell and Goldstein, 1998), maximizing the embedding distances between retrieval results while minimizing their distance to the claim. Ultimately, we output a subset of l diverse **sources** for the claim ($l < k$), augmenting each with its context before, after, and the text of its URL.

iii. Evidence & label generation module

1. We instruct a Large Language Model (LLM) to produce Question-Answer pairs required to fact-check given claim based on the provided sources, and predict its veracity verdict in a single output. We pass it the texts of all l sources, and several few-shot QA-pair generation examples picked from Averitec train set retrieved using BM25 based on the tested claim. The whole instruction is serialized into a system prompt and the format we used can be seen in Appendix A.
2. **Claim** is then passed to the LLM as a user message.
3. LLM is called to **generate the evidence** as a Question-Answer-Source triples and the Likert-scale scores for each possible **veracity verdict** in a single prediction, performing a chain of thought.
4. The LLM output is parsed, and the verdict with the highest score is chosen for the claim.

The main differences between this year's AIC FEVER 8 system, opposed to last year's AIC AVeriTeC system, are the omission of knowledge store pruning in the precomputation step[1], and, importantly, the choice of LLM.

2.1 Model and parameter choices

To produce our submission in the FEVER 8 shared task, the following choices were made to deploy the pipeline from section 2:

`mxbai-embed-large-v1` (Li and Li, 2024; Lee et al., 2024) is used for the vector embeddings, and the maximum chunk size is set to 2048 characters, considering its input size of 512 tokens and a rule-of-thumb coeffitient of 4 characters per token to exploit the full embedding input size and produce the smallest possible vector store size without neglecting a significant proportion of knowledge store text.

`FAISS` (Douze et al., 2024; Johnson et al., 2019) index is used as the vector database engine, due to its simplicity of usage, exact search feature and quick retrieval times (sub-second for a single FEVER 8 test claim).

$l = 10, k = 40, \lambda = 0.75$ are the parameters we use for the MMR reranking, meaning that 40 chunks are retrieved, 10 sources are yielded after MMR-diversification, and the tradeoff between their similarity to the claim and their diversity is 3:1 in favour of the source similarity to the claim (explained in more detail in Ullrich et al. 2024).

`Ollama` wrapper around `llama.cpp` is the LLM engine we use to deploy LLMs within the FEVER 8 test environment due to its robustness and ease of deployment.

[1]The precomputed vector stores were required to be independent on claim text in FEVER 8.

`Qwen3-14b` (Yang et al., 2025) is the LLM we use to produce the evidence and labels, we also let it generate its own `<think>` sequences, although further experimentation (Table 2) suggests that the thinking tokens may not justify the costs of their prediction, as they seem to perform on par with using only the evidence & label LLM outputs for its chain of thought.

3 Results and analysis

System	old AVeriTeC score	Q only (Ev^2R)	Q + A (Ev^2R)	new AVeriTeC score	time per claim
AIC CTU	0.41	0.20	**0.48**	**0.33**	54*s*
HUMANE	0.45	0.19	0.43	0.27	29*s*
yellow flash	0.16	0.16	0.41	0.25	32*s*
FZIGOT	0.46	**0.36**	0.40	0.24	19*s*
EFC	0.49	0.13	0.35	0.20	**7*s***
checkmate	0.38	0.18	0.34	0.20	22*s*
Baseline	**0.50**	0.27	0.34	0.20	34*s*

Table 1: FEVER 8 shared task system leaderboard as shared by organizers, listing new Ev^2R-recall-based (Akhtar et al., 2024) and legacy hu-METEOR AVeriTeC scores. Evaluated using AVeriTeC 2025 test set. Best scores are bold.

In Table 1, we reprint the final test-leaderboard of FEVER 8 shared task as provided by the organizers. Our system introduced in Section 2 scores first in the decisive metric for the task – the new AVeriTeC score – with a significant margin. This came as a surprise to its authors, as neither the values of the old, hu-METEOR-based AVeriTeC score (Schlichtkrull et al., 2024b), nor the dev-leaderboard available during system development phase (where our system scored 4th), suggested its supremacy. Let us therefore proceed with a discussion of possible strengths that could have given our system an edge in verifying the FEVER 8 test-set of previously unseen 1000 claims.

3.1 Why does the system perform well?

So why should our system outperform the FEVER 8 baseline and even the other systems submitted to FEVER 8 shared task despite the simplicity of its design (Figure 1) which boils down to a straightforward case of retrieval-augmented generation (RAG)?

The main reason, in our experience, is the large **context size** we opt for – while even the FEVER 8 baseline processes the claims and sources in a manner more sophisticated than we do, it processes the knowledge store on a *sentence* level, reducing the amount of information passed to the LLM as opposed to working with *documents* as a whole, which is the strategy our system approximates.

Despite our proposed integration of LLM into the pipeline being rather vanilla, combining sources of total length of as much as 60K characters[2] on model input yields highly competitive results, leveraging its own trained mechanisms of context processing.

Our other advantages may have been using a very recent model, Qwen3 (Yang et al., 2025), which naturally has a slightly higher leakage of 2025 claims into its train set than older models, and outperforms the previous LLM generations at long sequence processing. Furthermore, our pipeline design only uses a single LLM call per claim, meaning we could use the generously-sized 14B variant of Qwen3 and still match the time limit with Nvidia A10 and 23GB VRAM.

3.2 Scoring change impact

While the new AVeriTeC score based on Ev^2R-recall (Akhtar et al., 2024) estimates the proportion of correctly fact-checked claims[3] in all claims, just like the old hu-METEOR-based AVeriTeC score did, their underlying methods differ. Most importantly, an LLM-as-a-judge approach is now used instead of symbolic evidence comparison method. The rise of our system from 3rd place in AVeriTeC shared task (Schlichtkrull et al., 2024a) to 1st place in FEVER 8 without any major system change[4] can therefore also be attributed to the used scoring method. The old scoring method was, for example, found to be prone to some level of noise, as it was not robust against evidence duplication (Malon, 2024), which was a found exploit to boost evidence recall.

The discrepancy between old and new AVeriTeC score in Table 1 could motivate a further study on how the new score behaves, for example using

[2] In other words, around 33 standard pages. This number follows from our parameter choices in Section 2.1: 10 sources are retrieved for each claim, each with $\sim$ 2048 characters of the embedded text, and additional $\sim$ 4096 characters of context.

[3] Claims with sound evidence w.r.t. human annotation, and an exact match in predicted label.

[4] Despite scaling down.

the test-prediction files from last year AVeriTeC shared task systems. The familiarity of the systems, the availability of their hu-METEOR scores and documentation, may reveal valuable insights into the Ev^{2}R evaluation method itself, as in which behaviours does it punish and reward.

3.3 LLM impact

LLM	**Q only** (Ev^{2}R)	**Q + A** (Ev^{2}R)	***new AVeriTeC score***
GPT-4o$_{\texttt{2024-05-13}}$	0.30	0.58	0.40
Llama3.1-70B	0.37	0.54	0.39
qwen3:14B$_{\texttt{/no_think}}$	0.29	0.59	0.41
qwen3:14B$_{\texttt{/think}}$	0.20	0.59	0.42

Table 2: Ablation study on LLM choice and <think>-tokens impact on FEVER 8 dev-score. Pipeline design (Figure 1), retrieval results, system and user prompts are fixed. Evaluated using an on-premise Ev^{2}R scorer with Ollama-hosted Llama3.3-70B as a judge.

In 2024, we have experimented with then available versions of GPT-4o and Llama3.1-70B and found the open-source Llama to perform encouragingly well, depite the still-quite-cumbersome model size and the need for its quantization (Ullrich et al., 2024). This year, we have simply gone for the most recent open-weights LLM at the largest parameter count we could fit within our FEVER 8 compute budget, thus choosing the Qwen3 at its 14B parameter size (Yang et al., 2025).

Qwen3 was trained to produce thinking tokens by default, an approach popularized by DeepSeek (DeepSeek-AI et al., 2025) and OpenAI research models, to force the chain of thought. We have experimented with enabling and disabling this feature to see if it has an impact on the AVeriTeC score, and compared the model output quality to our last year prediction dumps, with evaluation experiments listed in Table 2.

Both Qwen3 evidence and label generation settings perform on par with previous GPT-4o generation, which validates our model choice. The thinking tokens, while producing legitimate-looking writeups of the fact-checking workflows (see Appendix B) were not shown to stimulate an improvement in AVeriTeC score in the ablation study (Table 2), so we suggest to disable this feature in future reproductions in favour of a faster prediction time (54s in the Table 1 was produced with the thinking feature *enabled*, so disabling it might solve the issue with near-limit runtime our pipeline suffers from).

4 Conclusion

In this paper, we have introduced our simple yet efficient RAG system which performed competitively well under time and compute constraints in FEVER 8 shared task, in May 2025. We release the used code along with usage instructions for producing the FEVER 8 submission, vector stores needed for the pipeline to run and their build scripts at `https://github.com/heruberuto/FEVER-8-Shared-Task/` which is a fork of the FEVER 8 baseline repository.

We attribute our success mostly to the use of *document* rather than *sentence* level of retrieval granularity and an employment of a recent LLM at a size which utilizes the whole compute and time budget with only around 10% time reserve as a failsafe. We encourage further usage of our system as a strong and easy-to-setup baseline for further research in automated fact checking and will be happy to answer any questions on the referred contacts.

4.1 Future works

1. Integrate a live search API as in (Malon, 2024) as a retriever into the AIC pipeline (Figure 1) to achieve a real-world generalization

2. Section 3.2 suggests to look at the key differences between legacy and Ev^{2}R scoring methods in terms of the available 2024 AVeriTeC leaderboard and available model documentations – we believe this could reveal valuable hints both scoring and pipelinne improvements in future work

Limitations

Our pipeline is not meant to be relied upon nor to replace a human fact-checker, but rather to assist an informed user. It gives sources and proposed labels for further questioning. It is optimized only for English, the carbon costs of the used models are considerable, despite the system trying to cut down the environmental cost of the prediction step.

Ethics statement

Our pipeline is an extension of our already existing last year submission all original authors agreed with, including the reusal of the necessary listing in Appendix A. The system was build specifically for the FEVER 8 shared task and reflects the biases of its annotators, for more information on this, we suggest the original AVeriTeC paper (Schlichtkrull et al., 2024b).

Acknowledgements

This article was created with the state support of the Ministry of Industry and Trade of the Czech Republic, project no. Z220312000000, within the National Recovery Plan Programme. The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged.

References

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2r: Evaluating evidence retrieval in automated fact-checking.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, ..., and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Malon. 2024. Multi-hop evidence pursuit meets the web: Team papelo at FEVER 2024. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 27–36, Miami, Florida, USA. Association for Computational Linguistics.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. InFact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112, Miami, Florida, USA. Association for Computational Linguistics.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024a. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024b. Averitec: a dataset for real-world claim verification with evidence from the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 137–150, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu ... Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.

A System prompt

```
You are a professional fact checker, formulate up to 10 questions that cover all
  the facts needed to validate whether the factual statement (in User message) is
  true, false, uncertain or a matter of opinion. Each question has one of four
  answer types: Boolean, Extractive, Abstractive and Unanswerable using the
  provided sources.
After formulating Your questions and their answers using the provided sources, You
  evaluate the possible veracity verdicts (Supported claim, Refuted claim, Not
  enough evidence, or Conflicting evidence/Cherrypicking) given your claim and
  evidence on a Likert scale (1 - Strongly disagree, 2 - Disagree, 3 - Neutral, 4 -
  Agree, 5 - Strongly agree). Ultimately, you note the single likeliest veracity
  verdict according to your best knowledge.
The facts must be coming from these sources, please refer them using assigned IDs:
---
## Source ID: 1 [url]
[context before]
[page content]
[context after]
...

---
## Output formatting
Please, you MUST only print the output in the following output format:
```json
{
 "questions":
     [
         {"question": "<Your first question>", "answer": "<The answer to the Your
           first question>", "source": "<Single numeric source ID backing the
           answer for Your first question>", "answer_type":"<The type of first
           answer>"},
         {"question": "<Your second question>", "answer": "<The answer to the Your
           second question>", "source": "<Single numeric Source ID backing the
           answer for Your second question>", "answer_type":"<The type of second
           answer>"}
     ],
 "claim_veracity": {
     "Supported": "<Likert-scale rating of how much You agree with the 'Supported'
       veracity classification>",
     "Refuted": "<Likert-scale rating of how much You agree with the 'Refuted'
       veracity classification>",
     "Not Enough Evidence": "<Likert-scale rating of how much You agree with the
       'Not Enough Evidence' veracity classification>",
     "Conflicting Evidence/Cherrypicking": "<Likert-scale rating of how much You
       agree with the 'Conflicting Evidence/Cherrypicking' veracity classification>"
 },
 "veracity_verdict": "<The suggested veracity classification for the claim>"
}
```

Few-shot learning
You have access to the following few-shot learning examples for questions and
  answers.:

Question examples for claim "{example["claim"]}" (verdict
  {example["gold_label"]})
"question": "{question}", "answer": "{answer}", "answer_type": "{answer_type}"
...
```

Listing 1: System prompt for the LLMs, FEVER 8 claim is to be entered into the user prompt. Three dots represent omitted repeating parts of the prompt. Reused and reprinted from Ullrich et al. 2024.

B Qwen3 thinking example

```
<think>
Okay, let's tackle this query. The user is claiming that due to Imran Khan's
  criticism of Macron's comments on Islam, French authorities cancelled the visas
  of 183 Pakistani citizens and deported 118 from the country. I need to verify
  this using the provided sources.

First, I'll look through the sources to see if there's any mention of visa
  cancellations or deportations related to Imran Khan's criticism.

Looking at Source ID 1, it talks about Imran Khan responding to Macron's remarks
  and the condemnation of Islamophobic policies, but there's no mention of visa
  cancellations or deportations.

Source ID 2 discusses a diplomatic row where Pakistan refused to accept a deportee,
  Muhammad Mansha, until his credentials were verified. This seems related to a
  specific case but not the large numbers mentioned in the claim.

Source ID 3 mentions protests in Pakistan against France's republishing of
  cartoons, but again, no visa or deportation numbers here.

Source ID 4 and 5 discuss diplomatic tensions, but not specific numbers of
  deportations. Source 5 and 6 from Newscomworld.com talk about fears of
  deportations of Pakistanis in France but don't mention specific numbers like 183
  or 118.

Source ID 7 is about Pakistan deporting French nationals, which is the reverse of
  the claim.

Source ID 10 mentions Pakistan summoning the French envoy but again, no numbers on
  deportations.

None of the sources mention 183 visas cancelled or 118 deported. The closest is
  Source 2, which mentions one individual, Muhammad Mansha, but that's a single
  case. The other sources discuss diplomatic tensions, protests, and possible fears
  of deportations but no actual numbers provided.

Therefore, the claim about 183 and 118 seems to be unsupported. The answer would be
  that there's no evidence of such numbers in the sources provided. The veracity
  verdict would be Refuted, as the sources don't support the specific numbers given.
</think>
```

Listing 2: Qwen3 thinking tokens example.