Implicit Context Compression for Local Software Engineering Agents

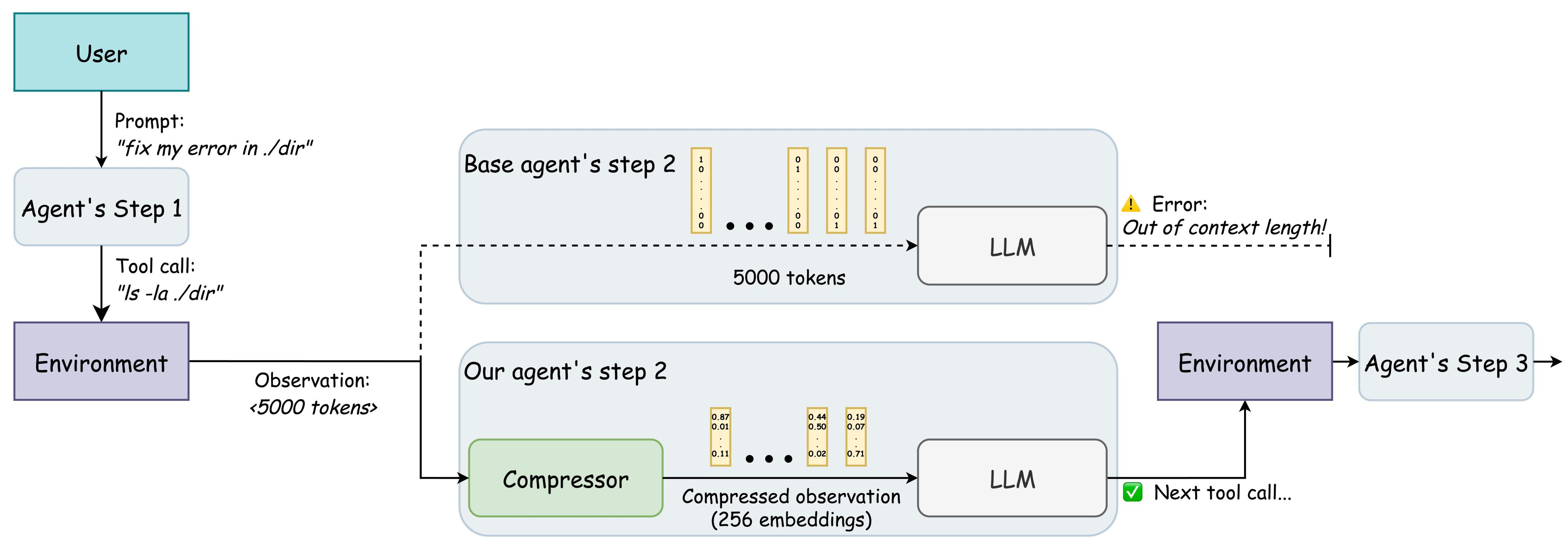
Kirill Gelvan*, Igor Slinko, Felix Steinbauer, Egor Bogomolov, Yaroslav Zharov

Motivation & Task Setup

Why and how?

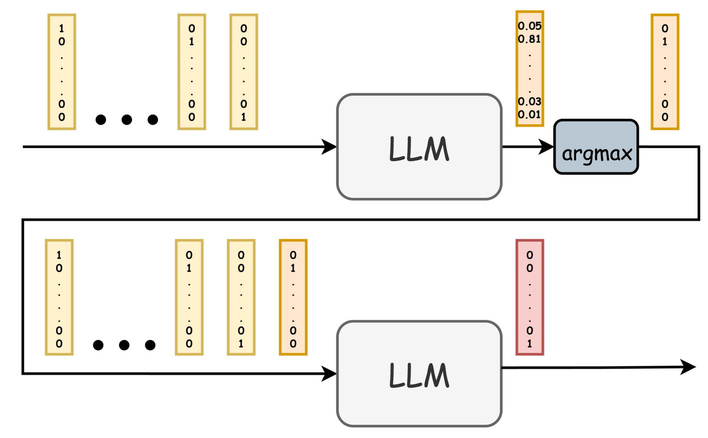
- LLMs have **limited context length**
- SWE Agents often call for tools with **unnecessarily long outputs**
- The **latent space** of embeddings is **much denser** than the discrete space of tokens
  ⟹
- Let's learn to compress tool outputs and use only the required information for the subsequent steps

SWE-bench Verified[1] example of Implicit Context Compression



Results of Implicit Compression without Training

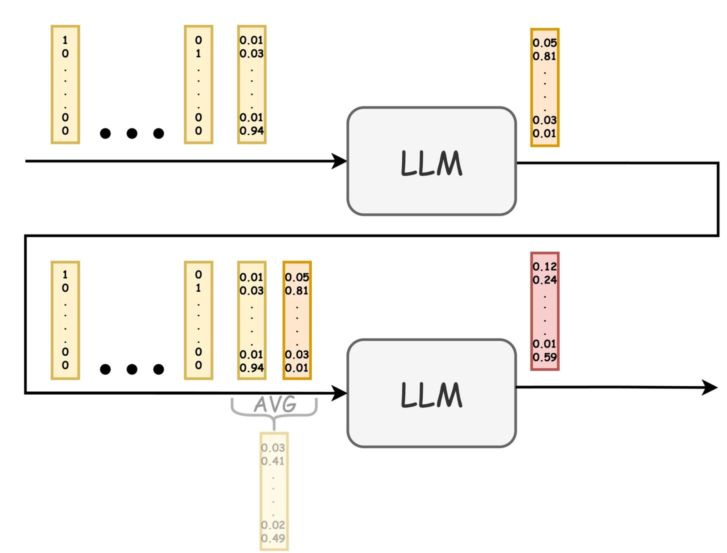
Hard Tokens (usual)



SQuAD[2], context embed	F1
Baseline — hard tokens	0.71

Table: Hard tokens technique

Soft Embeddings Tokens (ours)



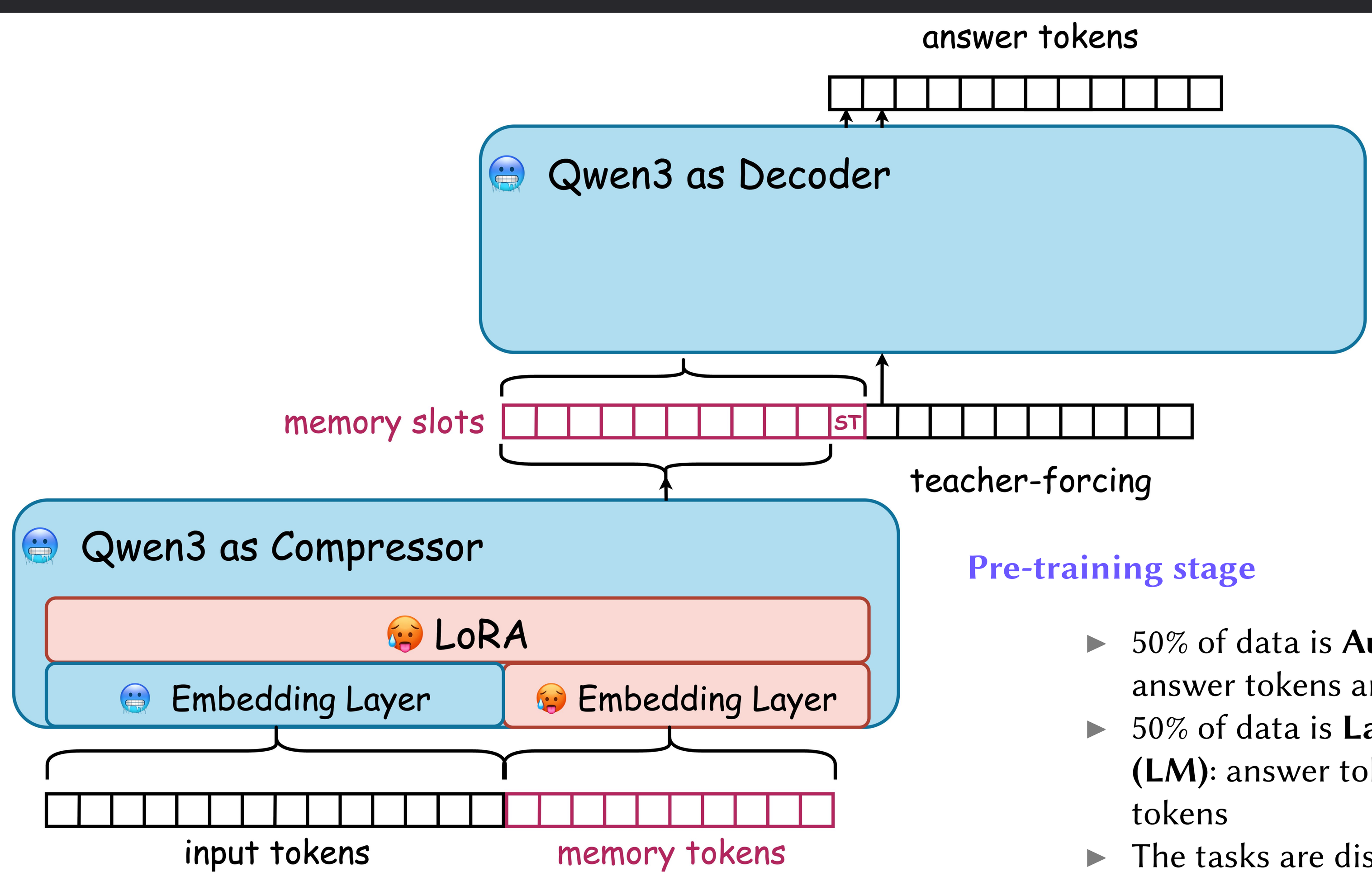
SQuAD[2], context embed	F1
Soft-embedded online	0.17
Soft-embedded online, avg ×2	0.11
Soft-embedded "regenerate-llm", avg ×2	0.16

Table: Soft-embedded techniques

Can we do it without training?

- No, we cannot!
- The scores drop dramatically (>50%) just by using continuous representations (without avg)
- In order to get to the latent space, **we need training**

ICAE – In-Context AutoEncoder [3]



Pre-training stage

- 50% of data is **AutoEncoding (AE)**: answer tokens are input tokens
- 50% of data is **Language Modeling (LM)**: answer tokens are continuation tokens
- The tasks are distinguished by a **special token (ST)** inserted in the middle

Fine-Tuning stage

- 100% of the data is from your own task. You may add a prompt after memory tokens as well

Note that only the "compressor" is trained, while **generation** is done by an **untouched Qwen** model!

Results of ICAE on General Texts & QA

Can we decompress texts after pre-training?

Dataset	Model	BLEU
PWC [3]	Mistral-7B	99.1
	Llama-2-7B	99.5
SQuAD (ours)	Qwen3-8B	98.1

Table: Text decompression on PWC and SQuAD

Can we solve Question-Answering task after fine-tuning?

Model	Compression	Exact Match (%)	F1
Mistral-7B	×1	49	68
LoRA-FT Mistral-7B	×1	59	65
ICAE-FT Mistral-7B	×1.7 ± 0.7	**69**	**73**

Table: ICAE averaging on SQuAD

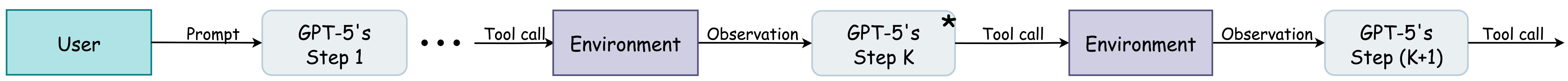
How can we interpret this?

- ICAE is able to decompress texts almost perfectly
- ICAE works with newer models and different datasets
- But does it work in an agentic setup?

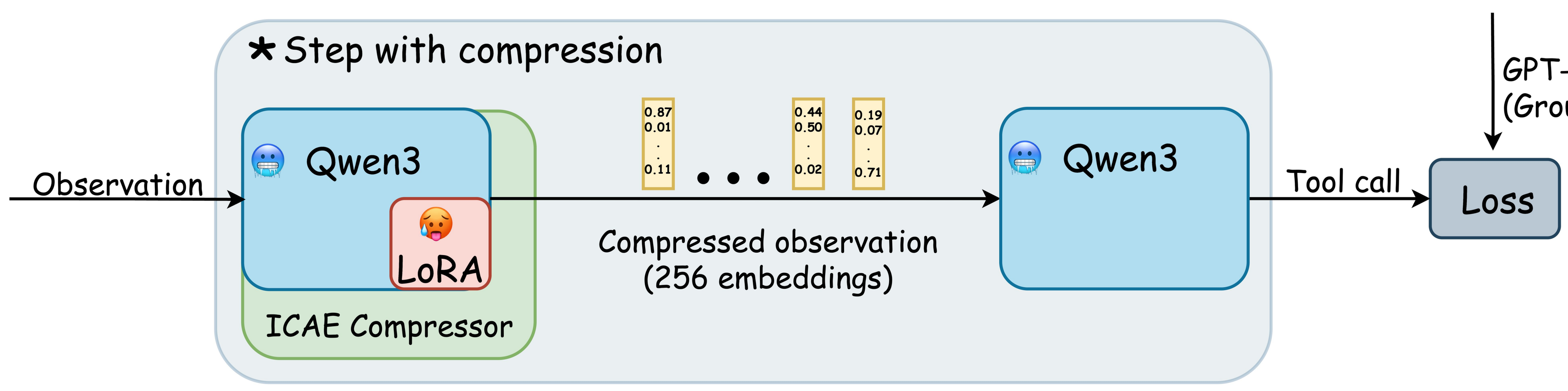
Training ICAE on SWE-bench

1 — Pre-train ICAE on general text corpora

2 — Get trajectories from a strong model (GPT-5)



3 — Fine-Tune Qwen3's LoRA on steps with compression:



Results

- **Latency**: ICAE compression shows a **10% faster** mean tool-call generation time than vanilla Qwen3-8B
- **Token-wise accuracy**: Qwen3-8B with and without compression perform **on par**
- **Resolved on SWE-bench Verified**: The model with compression **resolves fewer than 50%** of the original number of issues.

Hypotheses

- **Representation–behavior mismatch**: The ICAE encoder boosts token-level accuracy slightly, but perturbs decoder behavior for tool use, causing fewer end-to-end "resolved" completions
- **Compression trade-off**: Faster inference trims useful context or exploration, improving latency but reducing robustness on multi-step tasks required to count as "resolved"
- **Training dynamics / overfitting**: The low resolved count for the higher-accuracy variant suggests overfitting to labels rather than to execution reliability

[1] Neil Chowdhury et al. Introducing SWE-bench Verified. 2024.
[2] Pranav Rajpurkar et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text".
[3] Tao Ge et al. "In-context Autoencoder for Context Compression in a Large Language Model".

* Corresponding author: kirill.gelvan@jetbrains.com
JetBrains Research

More details here