Department of Mathematics
TUM School of Computation, Information and Technology
Technical University of Munich

Implicit Context Condensation for Local Software Engineering Agents

Kirill Gelvan

Thesis for the attainment of the academic degree

Master of Science

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Dr. Gjergji Kasneci

Advisors:
Felix Steinbauer
Igor Slinko

Submitted:
Munich, 31. November 2025

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 31. November 2025 Kirill Gelvan

Zusammenfassung

Eine kurze Zusammenfassung der Arbeit auf Deutsch.

Abstract

A brief abstract of this thesis in English.

Contents

1 Introduction

1.1 The Context Length Challenge in Large Language Models (LLMs)

The ability of Large Language Models (LLMs) to effectively process long sequences of input text is fundamentally constrained by their architecture. Specifically, Transformer-based LLMs face inherent limitations due to the self-attention mechanism, which scales quadratically with the number of tokens. Much previous research has attempted to tackle this long context issue through architectural innovations, but these efforts often struggle to overcome a notable decline in performance on long contexts despite reducing computation and memory complexity. While alternative attention mechanisms can make scaling closer to linear, they typically cannot achieve true linear scaling without quality drops on longer sequences.

The long context limitation presents a significant practical challenge, particularly in complex automated scenarios involving agents with many interaction turns. This restriction is worsened in software engineering (SWE) agent applications, where operational trajectories frequently involve tool calls that generate unnecessarily long outputs. SWE agents must perform tasks such as examining files and directories, reading and modifying parts of files, and navigating complex codebases. However, pretrained models literally cannot work with sequences longer than N (e.g., 32,000) tokens, which prevents them from efficiently processing the accumulated history generated by these tools.

Context compression offers a novel approach to addressing this issue, motivated by the observation that a text can be represented in different lengths in an LLM while conveying the same information. For example, the same information might be represented in various formats and densities without necessarily affecting the accuracy of the model's subsequent response.

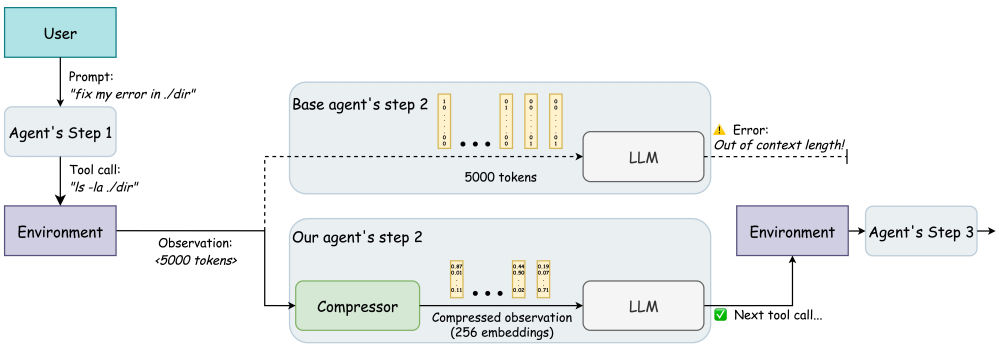

Figure 1.1 Comparison between base agent and our agent approaches for handling large observations exceeding LLM context length. The base agent fails when processing observations that are too long directly, while our agent successfully compresses the observation to 256 embeddings before LLM processing, enabling continued task execution.

The core goal of context condensation is precisely to leverage this potential density to enable LLM agents to execute tasks involving long chains of reasoning (or Chain-of-Thought, CoT) and more steps by condensing the environment observations. Achieving this condensation improves the model's capability to handle long contexts while offering tangible advantages in improved latency and reduced GPU memory cost during inference.

1.2 Implicit and Explicit Compression

The core concept of compression, motivated by the observation that information can be represented in different forms and densities, leads directly to the discussion and comparison between implicit and explicit approaches. Implicit compression moves beyond explicit, token-based techniques by utilizing the inherent density of continuous latent spaces.

Instead of relying on discrete representations, implicit compression focuses on mapping information into continuous representations (embeddings). This is possible because continuous latent spaces are inherently much denser than discrete spaces. Our goal is to enable more efficient processing of information by successfully leveraging these dense representations.

2 Background

2.1 Transformer Architecture and Positional Bias

Self-attention mechanism is well-known and described in detail in many articles. Thus, we focus on how positional signals influence which tokens are likely to interact, and why the layout of position identifiers (position IDs) matters for context compression.

Transformers are permutation-invariant over their inputs unless augmented with positional information. In the original formulation, absolute position encodings [Vas+17] (either fixed sinusoidal or learned) are added to token embeddings so that attention has access to token order. Subsequent works replace addition with position-dependent biases or transformations that make attention explicitly sensitive to token distances:

- Relative position representations add an index-dependent bias to the attention logits [SUV18].
- Rotary position embeddings (RoPE) apply a rotation to queries and keys so that their inner product becomes a function of relative displacement [Su+21].

Formally, for queries Q, keys K, and values V, attention often takes the form

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top + B}{\sqrt{d_k}}\right)V,$$

where B is a position-dependent bias. In relative schemes, $B_{ij} = b(i - j)$. In RoPE, the similarity $\langle Q_i, K_j \rangle$ implicitly depends on $i - j$ via rotations applied to Q_i and K_j.

These mechanisms create a local inductive bias: tokens that are closer in position tend to have higher prior attention affinity. This has direct implications for compression with special tokens ("memory", or compressed tokens): where those tokens are placed in position-ID space controls which parts of the sequence they can most easily interact with. Empirically, assigning position IDs to minimize distance between compressed tokens and the tokens they must interface with (either source content or the downstream prompt) improves effectiveness [Zha+25]. **Maybe say we would talk about it later in the thesis?**

2.2 Parameter-Efficient LLM Fine-Tuning

There are many techniques to efficiently fine-tune LLMs, but we focus on Low-Rank Adaptation (LoRA) [Hu+21]. LoRA freezes the pre-trained weight matrices and introduces trainable low-rank updates, yielding substantial parameter savings while preserving the base model's knowledge [Hu+21]. Consider a linear projection with base weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. LoRA parameterizes an additive update

$$\Delta W = BA, \quad A \in \mathbb{R}^{r \times d_{\text{in}}},\ B \in \mathbb{R}^{d_{\text{out}} \times r},\ r \ll \min(d_{\text{in}}, d_{\text{out}}),$$

so that the adapted layer computes

$$h = (W_0 + \alpha/r \cdot BA)\, x = W_0 x + \alpha/r \cdot B(Ax),$$

with a scalar scaling α controlling the update magnitude. Only A and B are trained; W_0 remains frozen. The trainable parameter count becomes $r(d_{\text{in}} + d_{\text{out}})$, dramatically less than $d_{\text{in}}\, d_{\text{out}}$ for typical ranks (e.g., tens to a few hundreds).

In Transformer blocks, LoRA adapters are commonly inserted in attention projections (query and value projections, see [Hu+21]) and optionally in output or feed-forward projections, trading off capacity and efficiency. The benefits include:

- parameter efficiency and reduced activation memory
- modularity—multiple task-specific adapters can be swapped atop a single base model
- faster fine-tuning

Practical considerations include choosing the rank r, the scaling α, as well as target layers to balance adaptation capacity and generalization [Hu+21].

2.3 Agentic Setup and Tool Use

We use "agentic" to refer to autonomous decision-making loops in which an LLM plans, invokes tools, and incorporates observations to pursue goals. At time t, the agent conditions on a history $H_t = [(a_1, o_1), \ldots, (a_{t-1}, o_{t-1})]$ and a task-specific system prompt s to select an action a_t. The environment (or tool) returns an observation o_t, which is appended to the history. This perception–action loop continues until termination. Concretely:

1. Prompt assembly: system instructions + task + concise guidelines + recent trajectory summary.
2. Policy step: the LLM proposes an action (often with brief rationale) and a tool to invoke.
3. Tool execution: the specified tool runs non-interactively with provided arguments.
4. Observation: tool output (stdout/stderr, structured JSON, file diffs, or retrieval snippets) is captured.
5. Memory update: (a_t, o_t) is logged; optional compression/summarization reduces footprint.
6. Termination or next step: the agent either returns a final answer or continues planning.

A prominent benchmark for evaluating such agentic capabilities is the Berkeley Function Calling Leaderboard (BFCL) [Pat+25]. BFCL provides a standardized suite of tasks to measure an LLM's proficiency in translating natural language requests into precise, executable tool calls. The tasks range from simple, single-function invocations to complex scenarios requiring multi-step reasoning and tool chaining. This benchmark exemplifies the practical challenges in agentic systems, where models must correctly interpret user intent and interact with external APIs or codebases [Pat+25].

Agentic toolcall examples (from [Pat+25]):

- Vehicle status lookup: `vehicle.getStatus(vin="WVWZZZ...")` — returns battery level, tire pressure, and last seen location.
- Driving route plan: `maps.route(origin="Munich", dest="Berlin", mode="driving")` — computes ETA and step-by-step directions.
- Hotel search: `booking.search(city="Prague", dates="2025-11-03..05", guests=2)` — lists options with price and rating.
- Weather check: `weather.current(city="Warsaw")` — returns temperature, precipitation, and alerts.

Code-oriented toolcall examples

- Code/bash execution: `execute(command="pytest -q")` or `execute(command="ls -la")` — runs unit tests using pytest or lists files with details.
- Symbol search: `find(name="parseUser")` — finds the definition and usages of a function in the codebase.
- String replacement in code: `str_replace(file_path="app.py", search="foo", replace="bar")` — replaces all occurrences of "foo" with "bar" in app.py.

3 Related Work

3.1 Architectural Approaches for Long Context Modeling

Long-context modeling must address the quadratic cost of vanilla self-attention [Vas+17] while preserving task-relevant dependencies over thousands of tokens.

A useful taxonomy distinguishes:

- attention variants that alter the attention operator itself;
- explicit compression methods that reduce the input via retrieval or summarization;
- implicit compression methods that learn compact, decoder-friendly representations without exposing full inputs during inference.

We briefly review each group and summarize advantages and limitations.

Attention variants. Sparse and local/windowed patterns reduce pairwise interactions to achieve subquadratic cost. Windowed attention restricts each token to a fixed neighborhood and optionally augments a small set of global tokens to propagate long-range information. Longformer combines sliding windows with learnable global tokens for long documents [BPC20]. Block- and mixed-sparsity patterns (e.g., banded + random) as in BigBird provide theoretical expressivity and empirical gains on long sequences [Zah+20]. Sparse Transformers [Chi+19] introduced fixed sparse patterns for scalable generation. Advantages include improved memory/computation and strong local modeling. Disadvantages include potential failures to route cross-window interactions when global tokens or connectivity patterns are insufficient, and hardware inefficiencies for irregular sparsity.

Linear-time approximations further change the attention operator. Kernelized attention (Transformers-as-RNNs) linearizes softmax attention for autoregressive decoding [Kat+20]. Linformer projects keys/values along sequence length to attain linear complexity [Wan+20]. These methods offer asymptotic gains and longer feasible contexts. However, they can underperform full attention on tasks requiring precise long-range interactions or exact softmax geometry. They also introduce approximation/projection hyperparameters that affect quality.

Explicit compression. Retrieval-augmented generation (RAG) or summarization-based pipelines reduce the effective input by selecting or rewriting content before decoding. RAG retrieves top-k passages from an external corpus and conditions generation on them. This approach improves knowledge-intensive tasks while decoupling parametric and non-parametric knowledge [Lew+20]. Abstractive summarization pre-compresses long inputs into concise proxies (e.g., PEGASUS pretraining with gap-sentence objectives) [Zha+20]. Benefits include controllable compute and access to external knowledge. Drawbacks include selection bias, retrieval latency, brittleness to retrieval errors, and potential loss of details critical for downstream reasoning.

Implicit compression. Here, the idea is to produce task-adapted representations (often embeddings) that a model uses during inference, rather than the input itself. Examples include learned soft/prefix prompts for steering frozen decoders [LARC21; LL21]. Other examples include tokenized memories trained to preserve answer-relevant information. Implicit methods maintain a tight interface to the model. They can reduce latency and memory without external retrieval. A key challenge of implicit compression is that, by condensing input into a compact intermediate representation, some information may inevitably be lost

and cannot be recovered with perfect fidelity. This may limit the utility of compressed representations, especially when critical details are omitted during the compression process.

Taken together, attention variants trade exactness for structure or approximation. Explicit compression trades completeness for selection. **TODO: this is bullshit but make a change to section 3./ we like implicit so we found ICAE—** — Implicit compression trades human readability for decoder-optimized compact interfaces.

3.2 Soft Prompting, Context Distillation, and Continuous-Thought Representations

Soft prompting is a method for conditioning large language models by learning continuous prompt vectors or prefix tokens rather than discrete text. These learned vectors are typically prepended to the model's input sequence and serve as a compact, trainable interface for adapting frozen decoders. Classic methods in this category include prefix-tuning [LL21] and prompt tuning [LARC21]. Both approaches involve optimizing a small set of continuous embeddings that steer the model toward desired behavior without full-scale model finetuning. This parameter-efficient interface enables flexible adaptation and can be tuned for domain- or task-specific goals. Because the prompt vectors are not constrained to map onto human-readable tokens, they can condense much more information than would be possible with standard textual prompts.

In practical, agentic settings the challenge of long or growing histories becomes acute (???).

CoConut (Chain of Continuous Thought) [**coconut_placeholder**; **arxiv_2412_06769**] generalizes the concept of soft prompting by moving away from natural language tokens altogether and enabling reasoning directly in latent, continuous spaces. Instead of relying on explicit tokenization and sequence rewriting, CoConut leverages the model's own hidden states as a "continuous thought" vector. After processing an input, the final hidden state (or a structured set of latent embeddings) is fed back as a contextual scaffold for further reasoning steps. Experiments show that directly reasoning in the model's own latent space improves downstream performance for tasks with extended, multi-step dependencies, outperforming classic chain-of-thought prompting. By discarding the constraints of discrete tokenization, CoConut demonstrates that agentic LLMs can reason, plan, and retain context in a fundamentally more expressive and compact way.

These three directions form a coherent spectrum of implicit compression strategies. All focus on designing interfaces for decoder-optimized, compact context that can scale to long, complex tasks without being forced back into the limits of surface-level, human-readable summaries.

3.3 The In-Context Autoencoder (ICAE) Framework

ICAE [Ge+24] is closely related to the above implicit compression paradigm. An encoder (often a LoRA-adapted copy of the base LLM) reads a long context and emits a small set of learnable *memory tokens*. A frozen decoder (the base LLM) then conditions on these tokens plus the downstream prompt to generate outputs. Training combines the following objectives:

- an autoencoding objective, prompting the decoder to reconstruct the original text from memory slots
- a language modeling/continuation objective that teaches the decoder to answer queries conditioned on the slots and a prompt

This design turns a long, potentially unwieldy context into a compact representation that the decoder can efficiently consume, improving latency and memory footprint while preserving fidelity for downstream tasks. The number of memory tokens controls the compression ratio (e.g., 4× or higher), and the position-ID placement of these tokens influences how readily the decoder can access stored information (cf. Section 2.1) [Ge+24].

Beyond the high-level description, Figure 3.1 depicts the encoder–decoder split. On the left, the encoder ingests the full context (e.g., a text) and produces a fixed number of memory tokens. On the right, the frozen decoder receives these tokens and a tokens During pretraining, the encoder is optimized so that the decoder can reconstruct the original text and continue language modeling (50/50 chance of being used). During fine-tuning, the objective emphasizes answering prompts correctly given only the memory tokens and the task prompt. In practice, the encoder is frequently adapted with parameter-efficient methods such as LoRA [Hu+21], whereas the decoder remains frozen to preserve the capabilities of the base model.

ICAE differs from heuristic summarization in that the representation is optimized for decoder consumption rather than human readability. It also differs from sparse or windowed attention in that the decoder still operates with dense attention over a small set of memory tokens. ICAE differs from explicit RAG in that no external retrieval is required at inference [BPC20; Lew+20; Zah+20]. Compared to purely recurrent memory, ICAE offers direct, content-dependent access via attention over a small set of tokens. This avoids long chains through recurrent states.

In the context of local software engineering agents, ICAE-style compression is particularly attractive: developer tooling yields verbose logs, test outputs, diffs, and compiler errors. Compact, learned memories can preserve salient facts (e.g., failing tests, stack traces, file paths, and prior fixes) while improving latency and GPU memory use. In later chapters we evaluate compression ratios around ~ 1.5–$2\times$ and analyze the accuracy–latency trade-offs compared to uncompressed baselines, as well as ablations that drop long or all observations to quantify the contribution of learned memory tokens.

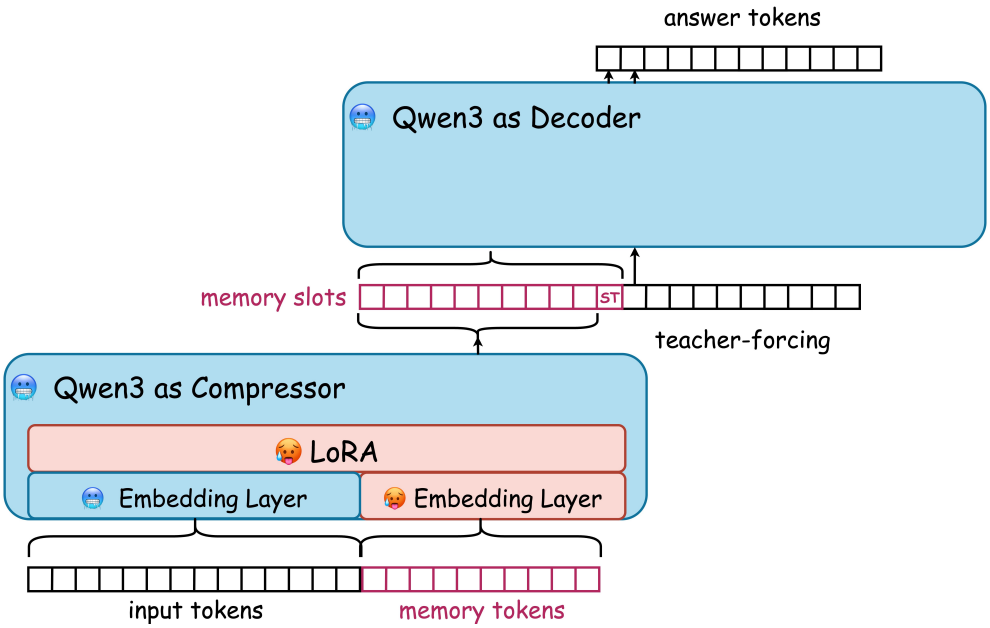

Figure 3.1 In-Context Autoencoder (ICAE) framework architecture

3.4 Some attempts to do the same thing?

Tobias' blogpost from openhands is not implicit, but explicit. so i have no knowledge of the same solutions using implicit compression?

4 Methods

4.1 ICAE Model Architecture and Components

The ICAE [Ge+24] structure consists of two modules: a lightweight encoder (LoRA-adapted LLM) and a fixed decoder (the target LLM, e.g., Qwen or Llama). The encoder processes the long context and appends learnable memory tokens (e.g., 128 tokens for 4x compression) to obtain memory slots. The decoder uses these memory slots, conditioned by prompts, to generate responses on behalf of the original context.

4.2 Self-Supervised Pretraining Objectives

Autoencoding (AE): The ICAE [Ge+24] is trained to restore the original input text from its memory slots, prompted by a special token [AE]. Language Modeling (LM) / Text Continuation: An additional objective where the model predicts the continuation of the context, prompted by a special token [LM]. This improves generalization and prevents overfitting to the AE task.

4.3 Instruction Fine-Tuning for Downstream Tasks

After pretraining, the ICAE [Ge+24] is fine-tuned using instruction data to enhance the interaction between the memory slots and various prompts, enabling the target LLM to produce desirable responses. The fine-tuning loss aims to maximize the probability of generating the correct response conditioned on the memory slots and the prompt.

4.4 Experimental Datasets and Configuration

General Text Data: The Pile dataset is used for pretraining. QA and Instruction Data: The PWC (Prompt-with-Context) dataset, consisting of thousands of (context, prompt, response) samples, is used for instruction fine-tuning and evaluation. SQuAD is used for offline quality measurement and fast debugging. Agentic Data: The SWE-bench Verified dataset is used to assess end-to-end performance on complex software engineering tasks, utilizing trajectories generated by a strong teacher model (e.g., GPT-5).

5 Experiments and Evaluation

5.1 Experimental Setup

We implement our ICAE framework from scratch, building upon the original architecture [Ge+24] with several modifications for improved efficiency and reproducibility. Our implementation uses Qwen3-8B as the base model, with LoRA adaptation applied to the attention matrices (q_proj and v_proj) using a rank of 128.

Pretraining is conducted on the SlimPajama-6B dataset using a combination of autoencoding and language modeling objectives, achieving 95% reconstruction BLEU score on general text. Fine-tuning on SWE-bench trajectories uses a larger memory size of 256 tokens and explicitly disables thinking mechanisms for simplicity, focusing on direct tool-call generation.

Training was performed on a single NVIDIA H200 GPU, requiring approximately 1 day and 15 hours for pretraining and 3 days for fine-tuning due to the computational complexity of the autoencoding objective and resulting lack of effective batching opportunities.

Detailed hyperparameters and training configurations are provided in Appendix A.1.

5.2 Initial Prototype Experiments: The Necessity of Training

The initial approach tested replacing hard tokens with soft/averaged continuous embeddings without fine-tuning. These prototype experiments, using methods like KV-cache hacks or direct embedding inputs in vLLM, demonstrated that scores decreased by more than 50% on QA tasks (e.g., SQuAD context embed F1 dropped from 0.71 to 0.17 or 0.11). This negative result confirmed the hypothesis that training is necessary to effectively condense context into the latent space.

Setting (SQuAD), context embed	Exact Match	F1
Baseline — hard tokens	**0.58**	**0.71**
Hard embedded, avg ×2	0.09	0.21
Soft embedded online, avg ×2	0.05	0.11
Soft embedded regenerate-llm avg ×2	0.07	0.16

Table 5.1 Baseline against averaging techniques (Prompt–Q–C)

5.3 Evaluation on General Text Reconstruction

Pretrained ICAE [Ge+24] demonstrated the ability to decompress general texts almost perfectly. High BLEU scores were achieved on datasets like PWC (99.1 for Mistral-7B, 99.5 for Llama-2-7B) and SQuAD (98.1 for Qwen3-8B), indicating that memory slots retained almost all context information for contexts up to 400 tokens. Analysis of reconstruction errors showed patterns similar to human memorization mistakes (e.g., restoring "large pretrained language model" as "large pretrained model"), suggesting the model selectively emphasizes or neglects information based on its understanding.

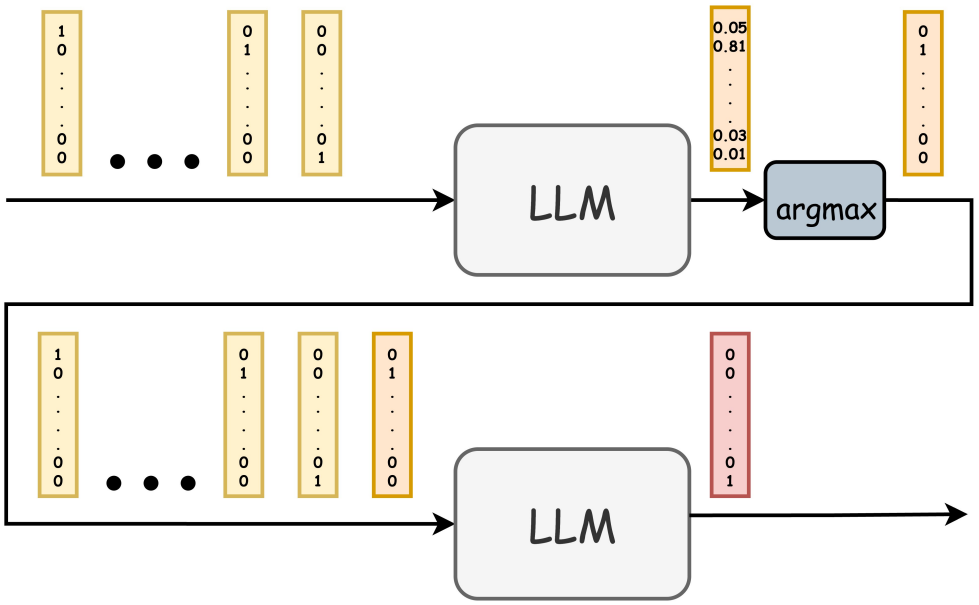

Figure 5.1 Results of the "without training" approach - SER1

5.4 Evaluation on Question Answering Tasks (Offline)

When fine-tuned on QA tasks (SQuAD), ICAE-FT [Ge+24] achieved high F1 (73) and Exact Match (69%) scores, performing well compared to LoRA-FT baselines. The quality of the compressed representation was shown to significantly outperform summaries generated by GPT-4 under the same length constraint (128 tokens).

Model	Compression	Exact Match	F1
Mistral-7B (no FT)	×1	49	68
LoRA-FT baseline	×1	59	65
ICAE FT (PwC, authors)	×1.7 ± 0.7	41	57
ICAE FT (SQuAD, ours)	×1.7 ± 0.7	**69**	**73**

Table 5.2 ICAE averaging on SQuAD

5.5 Evaluation on Agentic Performance (SWE-bench)

Efficiency Results: ICAE [Ge+24] compression led to measurable efficiency improvements, achieving a theoretically 10% faster mean tool-call generation time than the vanilla baseline (e.g., 0.4880s vs 0.5437s). Furthermore, latency tests showed speedups of 2.2× to 3.6× in total time for inference. Token-wise Accuracy vs. Resolved Rate: Although token-wise accuracy performed on par with (or slightly better than) the vanilla Qwen baseline (e.g., 0.9089 vs 0.9000), this metric was noted to be problematic ("token-wise accuracy is bullshit") and decoupled from true task success. End-to-End Task Success: The primary negative finding was that the model with compression resolved significantly fewer than 50% as many issues as the original Qwen model on the SWE-bench Verified dataset.

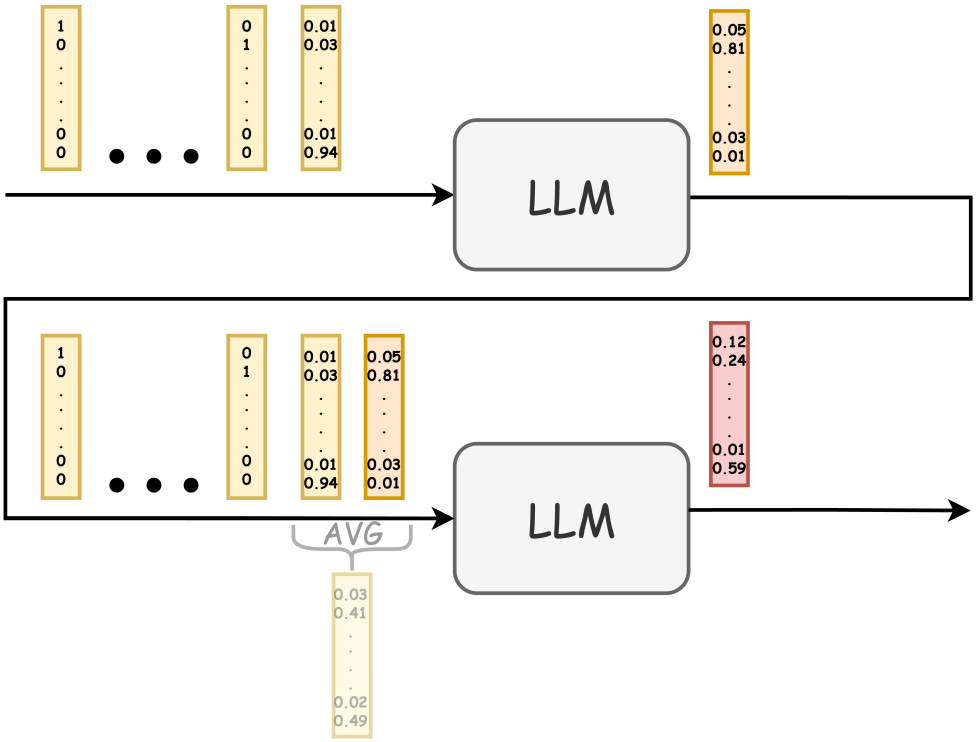

Figure 5.2 Results of the "without training" approach - SER2

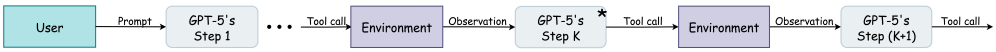

Figure 5.3 ICAE application to SWE-bench - Results 1

5.6 Discussion of Agentic Failure Hypotheses

Hypotheses for the end-to-end performance degradation include Representation–behavior mismatch, where the compression perturbs the decoder's behavior necessary for tool use. Other factors include reconstruction quality falloff for specialized content like code files, and potential overfitting to labels demonstrated by high local accuracy but low resolved rates.

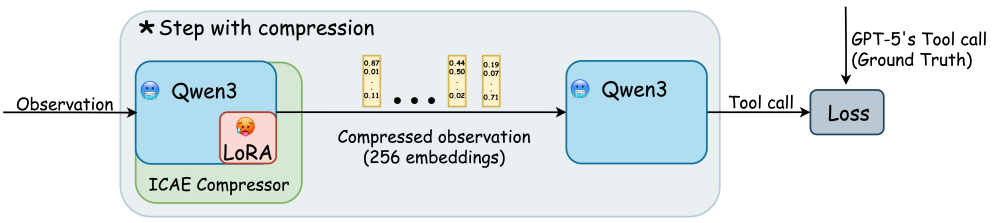

Figure 5.4 ICAE application to SWE-bench - Results 2

Encoder	Decoder	Accuracy	Mean tool-call time (s)
—	Full-FT	0.9484	1.24
—	LoRA-FT	0.9118	1.24
—	Qwen	0.8967	1.23
del long obs-s	Qwen	0.8873	0.44
del all obs-s	Qwen	0.8802	0.39
ICAE (LoRA-PT w/ Full-FT)	Full-FT	0.9219	—
ICAE (LoRA-PT w/ Qwen)	Qwen	0.8808	**1.12 (0.31+0.81)**
ICAE (LoRA-FT)	Full-FT	?	—
ICAE (LoRA-FT)	LoRA-FT	0.9263	—
ICAE (LoRA-FT)	Qwen	0.9020	—

Table 5.3 No think bug table. Qwen and ICAE future variants. FT=FineTuning, PT=PreTraining

Encoder	Decoder	Acc.	Time (s)	Resolved (/500)
—	Qwen-Full-FT	0.9484	1.24	—
—	Qwen-LoRA-FT	0.9118	1.24	10
—	Qwen	0.8967	1.23	26
ICAE (Qwen-LoRA-FT)	Qwen	0.9020	—	11
ICAE (Qwen-LoRA-FT)	Qwen-LoRA-FT	0.9263	—	3 (overfit?)
ICAE (Qwen-LoRA-FT)	Qwen-Full-FT	?	—	—
del long obs-s	Qwen	0.8873	0.44	1
del all obs-s	Qwen	0.8802	0.39	0
ICAE (Qwen-LoRA-PT w/ Q-Full-FT)	Qwen-Full-FT	0.9219	—	—
ICAE (Qwen-LoRA-PT w/ Qwen)	Qwen	0.8808	**1.12 (0.31+0.81)**	—

Table 5.4 No think bug table. Qwen and ICAE future variants. FT=FineTuning, PT=PreTraining

6 Limitations and Future Work

6.1 Reframing the Goal: Feature Extraction vs. Compression

When developing a context condensation strategy, the definition of success must be carefully framed. The approach investigated utilizes the In-context Autoencoder (ICAE) [Ge+24], which leverages the power of an LLM to compress a long context into short compact memory slots that can be directly conditioned upon by the decoder LLM.

However, the approach should be redefined not merely as "compression," which often implies a robust, high-ratio data reduction, but rather as "summarization" or "fixed length feature extraction," due to a core methodological constraint. This constraint arises from the hardcoded, non-robust nature of the intended output length (e.g., aiming for 256 tokens in general discussion).

This reframing is critical because lossless compression typically struggles to achieve ratios exceeding 10×. Under high compression ratios (lossy compression), the assumption that "all tokens in the context are equally important"—an assumption intrinsically aligned with lossless autoencoding training—is violated. When the information carrier capacity is limited, the compression mechanism should ideally focus only on the most important tokens, which conflicts with the uniform coverage implied by fixed-length encoding.

The fundamental mechanism supporting this goal is the relative density of different representation spaces. The latent space of embeddings is "much denser than the discrete space of tokens," which is the underlying justification for learning context condensation. Therefore, the thesis investigates how condensing environment observations (which contain irrelevant or redundant information) into continuous representations (embeddings/memory slots) affects agent performance and efficiency when addressing the context length challenge.

6.2 Limitations of Fixed-Length Compression

The methodology assumes that "all tokens in the context are equally important," aligning intrinsically with lossless compression (autoencoding), which becomes problematic under high compression ratios (lossy compression). The non-robustness of the approach is constrained by the hardcoded number of memory tokens (e.g., 256 tokens in discussion), defining the limitation of the approach as fixed-length feature extraction. Experimental results confirm that improvement attenuates or fails at high compression ratios (e.g., beyond 15x or 31x).

6.3 Constraints on Model Scale

Due to computational limitations, experiments were mainly conducted on Llama models up to 13 billion parameters.

6.4 Outlook for Future Research

Future work should explore validating the ICAE [Ge+24] effectiveness on larger and stronger LLMs, as performance is expected to benefit more from more powerful target models. Potential extension to multimodal LLMs (images, video, audio) is suggested, as these modalities have greater compression potential.

6.4.1 Open Source Contributions and Reproducibility

To advance the field of context compression for software engineering agents, we release our complete implementation, including pretrained models achieving 95% reconstruction BLEU and fine-tuned models that outperform uncompressed baselines on SQuAD. Our comprehensive release includes all training configurations, hyperparameters, and experiment logs, enabling future researchers to reproduce our results and build upon this work.

The open-source nature of this contribution addresses the reproducibility crisis in machine learning research, providing both the tools and transparency necessary for scientific progress in context management for LLM agents. All code, model checkpoints, and experiment logs are available at the project repository, with full Weights & Biases experiment tracking for both pretraining and fine-tuning phases.

6.5 Caching would not ever work with our approach? Or would it?

7 Conclusion and Outlook

7.1 Summary of Achievements

The ICAE [Ge+24] framework successfully achieves context condensation/feature extraction for general text, demonstrating high reconstruction quality (BLEU $\approx$ 99%) and measurable efficiency gains (speedup 2× to 3.6×). The work provided insights into LLM memorization patterns, suggesting similarities to human memory encoding.

7.2 Synthesis of Findings

Despite achieving efficiency and local accuracy on agent trajectories, the performance degradation in end-to-end task completion (resolved issues) highlights the critical gap between local context compression quality and robust decision-making in complex agentic settings.

7.3 Positioning the Work

The findings position this work within the broader research efforts on LLM context management, emphasizing the experimental results concerning condensation effectiveness across different data types, and suggesting caution when applying general compression methods to fine-grained, critical agent behaviors (code and tool use).

A Appendix

A.1 Training Details and Hyperparameters

A.1.1 Pretraining Configuration

Parameter	Value
Base Model	Qwen3-8B
Dataset	SlimPajama-6B
Learning Rate	1×10^{-4}
Batch Size	1
Gradient Accumulation	8
Training Steps	≈100,000
Memory Size	256 tokens (4× compression)
LoRA Rank	128
LoRA Target Modules	q_proj, v_proj
Optimizer	AdamW
Warmup Steps	300
Hardware	1× NVIDIA H200 GPU
Training Time	≈1 day 15 hours

Table A.1 Pretraining hyperparameters and configuration

A.1.2 Fine-tuning Configuration

Parameter	Value
Base Model	Qwen3-8B
Dataset	SWE-bench trajectories
Learning Rate	5×10^{-5}
Batch Size	1
Gradient Accumulation	1
Training Steps	≈150,000
Memory Size	256 tokens (4× compression)
LoRA Rank	128
LoRA Target Modules	q_proj, v_proj
Optimizer	AdamW
Warmup Steps	250
Hardware	1× NVIDIA H200 GPU
Training Time	≈3 days

Table A.2 Fine-tuning hyperparameters and configuration

A.1.3 Reproducibility Resources

To ensure full reproducibility, we publish our complete implementation including:

- Complete ICAE framework for both pretraining and fine-tuning phases (`https://github.com/JetBrains-Research/icae`)
- Full Weights & Biases experiment logs for pretraining: `https://wandb.ai/kirili4ik/icae-pretraining`
- Full Weights & Biases experiment logs for fine-tuning: `https://wandb.ai/kirili4ik/icae-swebench-finet`
- Pretrained model checkpoints achieving 95% reconstruction BLEU: TODO
- Fine-tuned models that outperform uncompressed baselines on SQuAD: TODO

A.2 Profiling Setup and Latency Measurement

Technical details of the test machine and runtime configuration used for latency measurements.

A.3 Detailed Evaluation Tables

Comprehensive tables of model performance, including token-wise accuracy, mean tool-call time, and resolved issues for various ICAE [Ge+24] variants (e.g., Qwen-LoRA-FT, ICAE (Qwen-LoRA-FT) Qwen).

List of Figures

List of Tables

Bibliography

[Ano25] Anonymous. *AttentionRAG: Attention-Guided Context Pruning for RAG*. 2025. arXiv: 2503.10720 [cs.CL].

[Ano24] Anonymous. *Block-Attention for Efficient RAG*. 2024. arXiv: 2409.15355 [cs.CL].

[BPC20] I. Beltagy, M. E. Peters, and A. Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL].

[Chi+19] R. Child et al. *Generating Long Sequences with Sparse Transformers*. 2019. arXiv: 1904.10509 [cs.LG].

[Cho+21] K. Choromanski et al. *Rethinking Attention with Performers*. 2021. arXiv: 2009.14794 [cs.LG].

[Ge+24] T. Ge et al. *In-context Autoencoder for Context Compression in a Large Language Model*. arXiv:2307.06945 [cs]. May 2024.

[Hu+21] E. J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.LG].

[Kat+20] A. Katharopoulos et al. *Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*. 2020. arXiv: 2006.16236 [cs.LG].

[LARC21] B. Lester, R. Al-Rfou, and N. Constant. *The Power of Scale for Parameter-Efficient Prompt Tuning*. 2021. arXiv: 2104.08691 [cs.CL].

[Lew+20] P. Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP*. 2020. arXiv: 2005.11401 [cs.CL].

[LL21] X. L. Li and P. Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. 2021. arXiv: 2101.00190 [cs.CL].

[Nak+21] R. Nakano et al. *WebGPT: Browser-assisted question-answering with human feedback*. 2021. arXiv: 2112.09332 [cs.CL].

[Pat+25] S. G. Patil et al. "The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models". In: *Forty-second International Conference on Machine Learning*. 2025.

[Sch+23] T. Schick et al. *Toolformer: Language Models Can Teach Themselves to Use Tools*. 2023. arXiv: 2302.04761 [cs.CL].

[SUV18] P. Shaw, J. Uszkoreit, and A. Vaswani. "Self-Attention with Relative Position Representations". In: *NAACL-HLT*. 2018.

[Su+21] J. Su et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. 2021. arXiv: 2104.09864 [cs.CL].

[Vas+17] A. Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. 2017.

[Wan+20] S. Wang et al. *Linformer: Self-Attention with Linear Complexity*. 2020. arXiv: 2006.04768 [cs.LG].

[Yao+22] S. Yao et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. 2022. arXiv: 2210.03629 [cs.CL].

[Zah+20] M. Zaheer et al. *Big Bird: Transformers for Longer Sequences*. 2020. arXiv: 2007.14062 [cs.LG].

[Zha+20] J. Zhang et al. *PEGASUS: Pre-training with Extracted Gaps Sentences for Abstractive Summarization*. 2020. arXiv: 1912.08777 [cs.CL].

[Zha+25] R. Zhao et al. *Position IDs Matter: An Enhanced Position Layout for Efficient Context Compression in Large Language Models*. arXiv:2409.14364 [cs]. Sept. 2025.