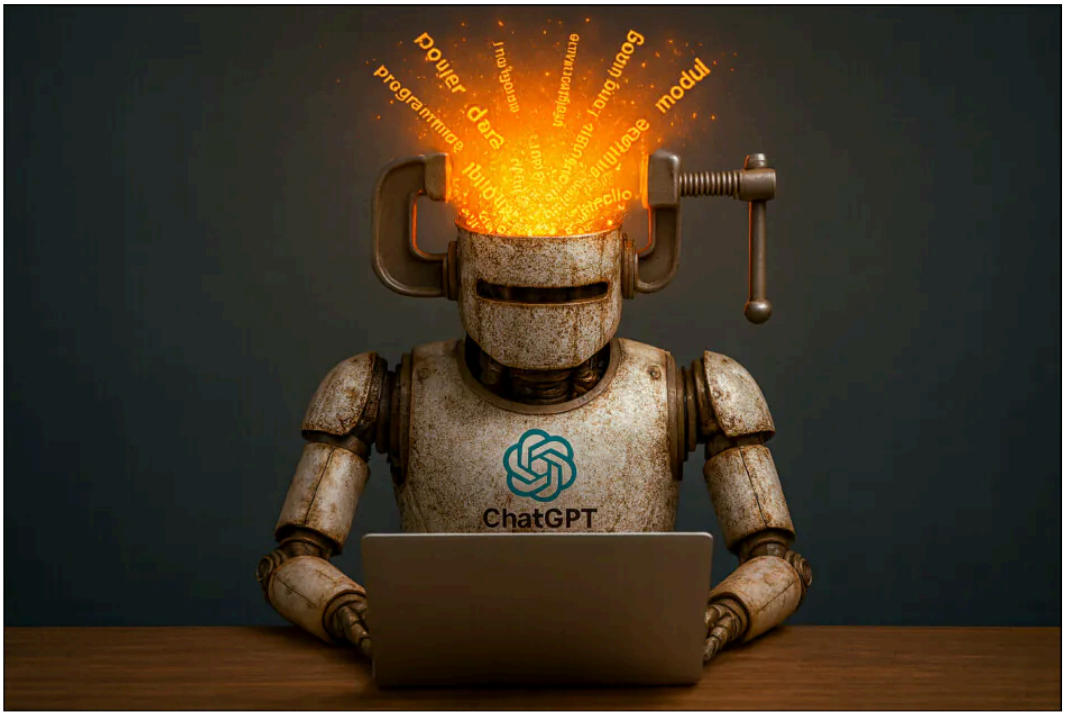
Teaching Forgetful AI To ‘Hold That Thought’ for Longer

Originally published at https://www.unite.ai/teaching-forgetful-ai-to-hold-that-thought-for-longer/



Language models often can’t remember the start of a conversation. A new text-compression method could change that, and make AI chat sessions much less infuriating.

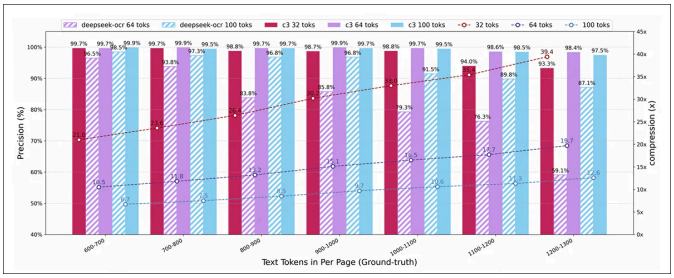
Conversational AI systems such as ChatGPT often lose track of earlier parts of a conversation, repeating themselves, or giving answers that ignore previously agreed rules.

This is because Large Language Models (LLMs) have a limited ability to focus, defined as a ‘context window’ of attention – like a torch that can light up only what it is directly aimed at, and a few adjacent objects.

Fixing such ‘amnesiac’ tendencies, which are due to these constraints upon attention, is among the most important directions of research into language-based AI models – not least because this syndrome severely limits the possibility of useful and consistent multi-turn conversations, and hamstrings the usefulness of LLMs in a variety of accuracy-critical contexts, such as medicine and law.

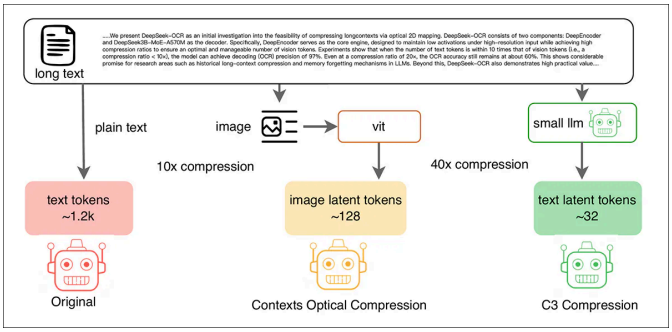
Crushing It

New research from China† is proposing a novel method of making a significantly larger amount of text fit into the limited resources of a GPU running an AI model – with results achieving a 20x compression improvement while retaining 98% accuracy:



Context Cascade Compression (C3) reconstructs long documents more accurately than optical compression methods such as DeepSeek-OCR, even when shrinking the input by up to forty times. Across a range of document lengths and compression settings, the new method maintains near-perfect fidelity, while the optical approach degrades sharply under higher compression. Source

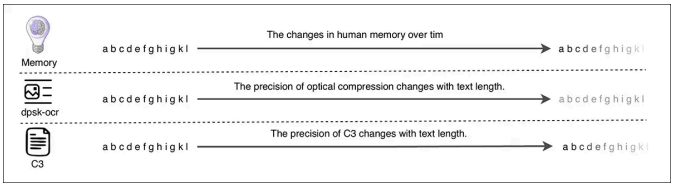
With an accuracy of 93% – which is within workable parameters – the text compression can even achieve a *40x* compression ratio:



Three approaches to compressing long text for language model input: the baseline method (left) tokenizes the text directly, yielding a large token count; the optical route (center) converts the text into an image and extracts visual embeddings using a Vision Transformer, achieving 10x compression; and the new C3 method (right) uses a small language model to compress the text into only 32 latent tokens, obtaining 40x compression without relying on visual encodings.

This means that the entirety of even a very long conversation can be compressed and re-injected (updated) at intervals into the exchanges as background context information, later in the chat – when the LLM would normally be forgetting earlier facts and slipping into 'amnesiac' behavior.

Though this is a lossy compression method, even the way that the loss occurs is useful: under the new method, the memory degrades at the *end* of a sentence, and not evenly throughout, as with the DeepSeek-OCR architecture that inspired the new approach; in fact, the researchers behind the new paper suggest that their method degrades in the same way as actual human memory, instead of randomly:



Top, human memory degrades at the end of the data stream; middle: DeepSeek-OCR degrades randomly, leaving no anchors that could aid in fixing the issue; bottom: the new method degrades in the same way as human memory, towards the termination of the data stream, offering markers that can help to improve accuracy through post facto *processing.*

This means that one can predict *where* the memorized data may be less reliable, and use this knowledge to address the problem – potentially offering a massive improvement in conversational recall and coherence, with 100% accuracy after remediation.

The new approach is called *Context Cascade Compression* (C3), and is inspired by the way that DeepSeek-OCR compresses text as images, achieving great compression levels. However, by using two (medium and large) language models to crunch long text directly down into latent embeddings, the new approach cuts out the drag caused by using raster images, thus achieving the improved performance.

The paper states:

> *'The superior performance of C3 can be attributed to its fundamental architectural design. The Deepseek-OCR analysis hypothesizes that its performance decline is due to factors like "complex layout" and "image blurring at lower resolutions"– inherent limitations of the optical pathway.*
>
> *'Our C3 paradigm, by operating directly in the textual domain, is entirely immune to these visual-domain artifacts. It avoids the information loss associated with rendering text to pixels and then encoding those pixels. Instead, it leverages a pre-trained LLM's powerful semantic understanding to distill textual information directly into an efficient latent representation.'*

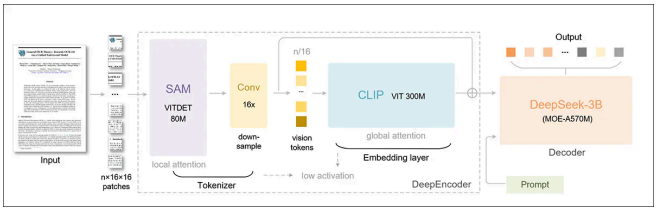
The new paper is titled *Context Cascade Compression: Exploring the Upper Limits of Text Compression*, and comes from two authors†, who appear* also to be offering C3 as an open source repository at GitHub.

Method

To understand the new approach, it's useful to know what optical character recognition (OCR) is, because this is where the whole idea stems from.

OCR is an algorithmic method dating back to the 1920s, though popularized in the 1990s, where pattern-detection allows a computer program to turn raster text (i.e., text inside images, that cannot be selected and which only exists as photographic content) into editable text.

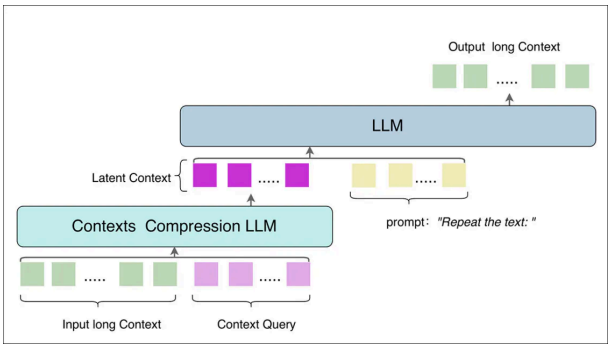
The creators of DeepSeek-OCR discovered that text could be compressed much more than standard pipelines by using OCR as an intermediary stage. In other words, instead of compressing text in itself, one could achieve a greater density of latent embeddings (i.e., more information saved) *by compressing a rasterized version of that text*:



From the DeepSeek-OCR release paper, a schema for the compression pipeline, including 16×16 rasterized patches as the OCR component. Source

The new paper suggests that optical approaches such as DeepSeek-OCR may be misattributing the source of their compression gains. Instead of the shift from text to image being the primary factor, the authors posit that the key benefit comes from converting verbose text tokens into more efficient latent representations.

To test this, they created a pipeline leveraging two language models: the smaller Qwen2.5 1.5B, which functions as the encoder, compressing long passages into a small set of latent tokens; and the larger Qwen2.5 3B, which functions as the decoder, reconstructing the original text from those tokens:



In the new C3 system, the smaller Qwen2.5 1.5B model compresses a long input into a fixed-length set of latent tokens, using trainable query embeddings. These tokens, together with a prompt, are passed to the larger Qwen2.5 3B model, which reconstructs the original text. This architecture enables high-fidelity recall of long sequences using only a fraction of the original token count.

To handle the compression stage, the researchers adapted the pre-trained Qwen2.5 1.5B model by introducing *trainable query embeddings*: abstract prompts that guide the model in distilling the incoming long context into a much smaller latent representation.

Rather than modifying the architecture, the method simply feeds the long text and the query embeddings together as a single input. The model's self-attention mechanism treats these elements identically, allowing it to output a fixed-length latent context without needing new layers or design changes; and this output is then handed off to the larger model for reconstruction.

To evaluate how much information survived compression, the researchers instructed the Qwen2.5 3B decoder to reconstruct the original input using only the latent tokens, and the prompt *'repeat the text'*. Since the task involved exact reproduction, rather than summarization or paraphrase, any deviation from the original could be directly traced to information lost in compression, offering a clean and objective test of fidelity across the encode-decode pipeline.

Data and Tests

The paper states that the authors compiled an original dataset of OCR material, amounting to a million pages obtained from the internet. No further detail is apparent on this sourcing, and the authors seem to be deliberately vague on this point.

Nonetheless, they observe that data engineering and curation was unnecessary for their purposes, and that they were able to train their model effectively on samples of 'diverse length'; and they state that this indicates their architecture to be resilient (and by inference, well-generalized, depending on training settings).

The model was trained on a high-performance cluster consisting of eight NVIDIA H800 GPUs, each equipped with 80GB of VRAM, for a total VRAM resource of 640GB. Each GPU accommodated a batch size of 2, obtaining a total global batch size of 256, when taking into account the 16 accumulation steps scheduled. The optimizer was AdamW, over a total of 40,000 steps.

To test the effectiveness of the C3 architecture, the researchers followed the same evaluation setup used in the original DeepSeek-OCR paper, using the Fox benchmark to measure compression and reconstruction accuracy across a range of document lengths.

English-language texts were selected, with passages ranging from 600 to 1300 tokens, with tokenization performed using the Qwen tokenizer.

To enable fair comparison, equivalent compression levels from the (DeepSeek-OCR) optical baseline were used, with 64 and 100 latent tokens. To explore the limits of the method, additional tests were run using only 32 latent tokens. In every case, reconstruction was initiated with the instruction *'repeat the text:'*:

	Latent Tokens = 64			Latent Tokens = 100		
Text Tokens	C3	DS-OCR	Compression	C3	DS-OCR	Compression
600-700	99.7%	96.5%	10.5×	99.9%	98.5%	6.7×
700-800	99.9%	93.8%	11.8×	99.5%	97.3%	7.5×
800-900	99.7%	83.8%	13.2×	99.7%	96.8%	8.5×
900-1000	99.9%	85.9%	15.1×	99.7%	96.8%	9.7×
1000-1100	99.7%	79.3%	16.5×	99.5%	91.5%	10.6×
1100-1200	98.6%	76.4%	17.7×	98.5%	89.8%	11.3×
1200-1300	98.4%	59.1%	19.7×	97.5%	87.1%	12.6×

Results from the initial test. Reconstruction precision and compression ratios were measured across seven token ranges using 64 and 100 latent tokens, showing consistent outperformance of C3 over DeepSeek-OCR, especially at higher compression levels.

Discussing the initial results visualized above, the paper states:

> *'The data unequivocally demonstrates that C3's direct text-to-latent compression paradigm significantly outperforms the optical compression approach across all tested conditions, establishing a new state-of-the-art in high-fidelity context compression.'*

When both systems were tested on longer documents, DeepSeek-OCR started to lose accuracy as compression increased, dropping below 60% in the most extreme case. C3 handled the same level of compression with much less loss, holding steady near 98%, even when the input was shrunk to a *twentieth* of its original size.

In the most demanding test, full texts were shrunk to just 32 tokens. Even then, the model managed to recover almost all the original content, keeping accuracy close to 99% in many cases:

Latent Tokens = 32		
Text Tokens	precision	Compression
600-700	99.7%	21.0×
700-800	99.7%	23.6×
800-900	98.8%	26.4×
900-1000	98.7%	30.2×
1000-1100	98.8%	31.0×
1100-1200	94.0%	35.4×
1200-1300	93.3%	39.4×

Reconstruction accuracy and compression ratios at 32 latent tokens, showing that even at extreme reductions (up to nearly 40x) precision remains above 93%.

At the most extreme setting, where the input was compressed to nearly one-fortieth its size, it still recalled over 93%. By comparison, earlier optical methods dropped to about 60% accuracy at half that level of compression.

The authors state††:

> *'These findings conclusively demonstrate the advantage of the C3 architecture. By avoiding the information bottlenecks and potential artifacts inherent in the visual modality (e.g., image resolution limits, layout complexity), our method gracefully handles extreme compression with minimal information loss.*
>
> *'The results from this aggressive test case solidify our claim that direct text-to-latent compression is a fundamentally more efficient and powerful paradigm than its optical counterparts.'*

They also conclude that C3 can *'unlock new capabilities in processing entire books, extensive legal documents, or large codebases for tasks like question-answering, summarization, and analysis'*.

Conclusion

This is one of the clearest and most approachable papers I've come across recently, with a commendably clear core idea that may prove to at least be an additional line of attack against the 'context window problem'.

Many users of LLMs will have learned by now to 'refresh' crucial information or guidelines periodically throughout a long exchange, having discovered the hard way that the likes of ChatGPT cannot retain that information for very long. Following on from this instinctive quick-fix, the idea at hand in the new paper is that a highly compressed version of a long conversation could be re-injected at intervals, automatically, into the context window of the current LLM instance, essentially 'remembering by proxy'.

In a climate where GPU scarcity is leading to a price rally even on traditional computer memory (such as DRAM, where many former GPU workloads are now being offloaded to support larger models), it seems unlikely that AI's host hardware is likely to become significantly more capacious in the near future; therefore innovative approaches such as this, which are almost a nostalgic rerun of the 1990s evolution of file compression, may prove necessary to drive performance forward.

More importantly, it would just be great if LLMs could maintain a coherent conversation for an hour or more, and actually remember what the conversation was about in the first place.

* *CLI installation instructions are given, but I am not able to take time to attempt an installation, and am not certain that the repo is code-complete.*

† *The two authors are stated as Fanfan Liu and Haibo Qiu. As far as casual research indicates, Qiu is currently a researcher at Chinese technology company Meituan, and Liu is an MS student at the Chinese Academy of Sciences. Neither currently has the paper in question listed in their histories. If either of these attributions are incorrect, please contact me via my profile.*

†† *Though the authors stick to the formula by providing additional qualitative tests, these are unilluminating compared to the earlier round of compression tests, and I have not covered them here.*

First published Friday, November 21, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai