Language Model-Assisted Text Compression and Decompression

…This is a draft idea of a technique that can be used to compress text keeping its semantic meaning…

The main idea

A method that leverages the capabilities of advanced language models to condense a text into a compressed, “primitive version”, and subsequently employs a separate language model to attempt the recovery of the original text's meaning. This two-step process aims not only to preserve the semantic integrity of the text during compression but also to efficiently regenerate a detailed approximation of the original content through decompression.

Experiment example

Original text fragment

Text	**Tokens**	**Words**	**Characters**
Tablets are no longer just personal devices for watching movies and catching up on emails in your downtime. Many can hold their own against laptops, especially in situations where portability and touchscreens are essential. Professionals working outside the office increasingly prefer them over bulkier laptops, as they can deliver the same level of productivity with a more flexible user experience. In an age of increasing hybrid and remote work, the use case of a highly portable tablet replacing a notebook is more of a consideration than ever before.	100	88	554

Compressed version (after applying the first step)

Text	**Tokens**	**Words**	**Characters**
Tablets not just for movies/emails. Now rival laptops; key for portability/touch. Preferred by remote workers over	*45* *(45%)*	34 (37%)	229 (41%)

big laptops for flexibility and productivity. In remote work era, tablets as notebook replacements more considered			

Recovered version (after applying the second step)

Text	**Tokens**	**Words**	**Characters**
Tablets have evolved significantly beyond their initial uses for watching movies and checking emails. Nowadays, they can compete with laptops in terms of functionality. One of their key advantages is portability, combined with touch screen capabilities, making them highly convenient for use on the go. They have become increasingly popular among remote workers who prefer them over larger laptops due to their flexibility and the productivity benefits they offer. In the era of remote work, the idea of using tablets as replacements for traditional notebooks is gaining more consideration, thanks to their lightweight design and versatile features that suit a variety of professional needs.	117	104	691

Similarity checking (using cosine distance)

Comparison	**Similarity**
Original vs Compressed	85.94%
Compressed vs Regenerated	88%
Original vs Regenerated	90.67%

LLM models

LLM model used: OpenAI GPT 3.5
Embedding model used: OpenAI text-embedding-3-small

Prompts

Text compression prompt

This is a draft of the prompt used to compress the original text. It contains the breaking down the instructions for the compression steps and a placeholder in the end where we inject the text to be compressed.

```
Apply the following methodology to compress the provided text while preserving its core meaning:

---
# Methodology: Contextual and Semantic Compression of Textual Content

This section delineates our systematic approach to achieving contextual and semantic compression of textual content. Our objective is to distill sentences to their essence, ensuring the preservation of their original meaning while minimizing verbosity. This methodology leverages foundational linguistic principles and a deep understanding of context to eliminate superfluous elements without compromising the integrity of the information conveyed. It finds particular utility in applications requiring succinct communication, such as text summarization, transmissions in bandwidth-limited environments, or cognitive processing tasks.

## Content Identification

### Objective
To ascertain and highlight the primary message(s) inherent within the text.

### Procedure
An initial analysis of the text is conducted to identify its central themes or ideas. This preliminary step is crucial for setting the stage for effective compression, as it identifies the indispensable elements that must be retained to preserve the text's original intent and significance.

### Example
Consider the sentence, "Tablets are no longer just personal devices for watching movies and catching up on emails in your downtime." The primary message distilled from this sentence is the expanded utility of tablets beyond mere leisure activities.

## Elimination of Redundant or Supplementary Information

### Objective
To excise words or phrases that do not contribute meaningfully to the primary message.

### Procedure
The sentence undergoes a rigorous examination to remove any adjectives, adverbs, and subordinate clauses that serve only to embellish the narrative without adding substantive information. This refinement focuses attention on the core ideas.
```

### Example
The phrase "in your downtime" is identified as non-essential and subsequently removed.

## Simplification of Vocabulary

### Objective
To replace complex or niche terms with simpler, more universally comprehensible alternatives.

### Procedure
Terms characterized by complexity, professional jargon, or specificity are substituted with more accessible language, ensuring the compressed text remains intelligible to a wider audience.

### Example
"Personal devices" is streamlined to "for," and "watching movies and catching up on emails" to "movies/emails."

## Reduction to Key Phrases

### Objective
To distill sentences to their fundamental phrases, capturing their essence.

### Procedure
Sentences are deconstructed into their basic components—subjects, verbs, and objects—with only those critical for conveying the overarching message retained. This process prioritizes succinctness and clarity.

### Example
The sentence is condensed to "Tablets not just for movies/emails."

## Use of Symbols/Abbreviations

### Objective
To employ symbols or abbreviations for succinct representation of common words or ideas.

### Procedure
Commonly understood symbols or abbreviations are utilized to replace standard terms, further condensing the text without sacrificing clarity.

### Example
The use of "/" effectively conveys "and" or "or," efficiently compressing the phrase.

## Reassembly into Coherent Phrases

### Objective
To logically organize the simplified components into a coherent structure.

### Procedure
The distilled elements are sequentially arranged to ensure the compressed message retains logical flow and comprehensibility.

```
### Example
"Tablets not just for movies/emails" exemplifies the outcome of this thoughtful reassembly.

## Final Review for Content Integrity

### Objective
To confirm that the compressed text faithfully retains the complete meaning of the original.

### Procedure
The final compressed version is meticulously compared against the original text to ensure all
critical information and its intended meaning are intact. Adjustments are made as necessary
to uphold the integrity of the content.

### Example
The rephrased sentence accurately communicates the expanded functionality of tablets,
demonstrating the compression's efficacy.

## Contextual Compression (Optional)

### Objective
To leverage implicit knowledge or shared context to omit explicit mentions of inferable
information.

### Procedure
The anticipated background knowledge of the target audience is evaluated to discern which
concepts might be considered common knowledge or easily inferred. Explicit references to
such concepts are then eliminated, relying on the audience's inferential capabilities to grasp
the compressed message within a shared context.

### Example
By presuming a mutual understanding of tablets' multifunctional capabilities, the phrase
"Tablets for more than leisure" succinctly encapsulates the core message.
---

Your task is to apply this methodology to the following text and return only the final
compressed version:

<<{input_text}>>
```

Text expansion prompt

This is a draft for the text expansion prompt. It also contains a placeholder in the end of the prompt that expects the text to be expanded.

```
Apply the following methodology to expand the provided text:
```

# Methodology: Semantic Expansion of Compressed Textual Content

This section presents our methodology for the semantic expansion of compressed textual content, aimed at reconstructing the original or an equivalently detailed version of a text from its condensed form. The process emphasizes maintaining the semantic integrity and informational depth of the original content while expanding the compressed text. This methodology is vital for applications such as detailed document synthesis, enhancing comprehension, or facilitating deeper analysis of condensed texts.

## Understanding Compressed Content

### Objective
To interpret and comprehend the core meanings encapsulated within the compressed text.

### Procedure
Analyze the compressed text to discern its key messages, themes, and any abbreviated terms or symbols. This foundational understanding is critical for accurate expansion that aligns with the original text's intent.

### Example
Given the compressed phrase "Tablets not just for movies/emails," the interpretation recognizes tablets' multifunctional use beyond entertainment and communication.

## Reintroduction of Contextual Details

### Objective
To restore the specific details and contextual nuances that were omitted during compression.

### Procedure
Infer and reintegrate contextual elements relevant to the compressed content's subject matter, ensuring these additions enhance the text's clarity and depth without deviating from the original message.

### Example
Expanding "Tablets not just for movies/emails" might involve specifying alternative uses of tablets, such as "Tablets are used for a wide range of activities beyond watching movies and managing emails, including..."

## Vocabulary Enrichment

### Objective
To enhance the text's richness and precision by incorporating a diverse vocabulary.

### Procedure
Identify areas within the expanded text where the language can be diversified or specified to better convey the nuances of the original message. Replace or augment simple terms with more descriptive language or specific jargon relevant to the text's context.

### Example
For the expanded content regarding tablets, "used for a wide range of activities" can be enriched to "serve diverse functions from educational purposes to professional productivity, illustrating their versatility beyond mere entertainment and communication."

## Reconstruction of Complex Sentences

### Objective
To rebuild the syntactic complexity of the original text, reflecting its structural depth and linguistic nuances.

### Procedure
Transform simplified sentence structures into more complex forms by reintroducing subordinate clauses, varied sentence types, and transitional phrases. This step ensures the expanded text mirrors the original's sophistication and style.

### Example
The basic sentence "Tablets serve diverse functions" can be reconstructed to "Tablets, once primarily seen as devices for leisure, have now emerged as versatile tools that serve a myriad of functions, ranging from education to professional productivity."

## Integration of Supplementary Information

### Objective
To incorporate additional information that supports or enhances the original message, providing a more comprehensive understanding.

### Procedure
Supplement the text with relevant data, examples, or explanations that were not present in the compressed version but are beneficial for achieving a fuller comprehension of the topic.

### Example
Adding to the tablet's expanded description, one might include "Recent advancements in tablet technology have further extended their capabilities, allowing for sophisticated applications such as graphic design, augmented reality experiences, and high-level computational tasks."

## Review for Semantic Fidelity

### Objective
To ensure that the expanded text faithfully represents the original content's meaning, context, and intent.

### Procedure
Compare the expanded text against the original (if available) or the intended message of the compressed version, making adjustments as necessary to maintain accuracy and completeness of information.

### Example

```
Reviewing the expanded tablet text to ensure it accurately reflects the evolution and
multifunctionality of tablets without introducing misinterpretations or inaccuracies.

## Final Editing for Coherence and Flow

### Objective
To refine the expanded text for readability, ensuring it is logically structured and flows
smoothly from one idea to the next.

### Procedure
Perform a final edit to adjust sentence transitions, coherence, and overall structure, ensuring
the text is engaging and easy to follow.

### Example
Ensuring the expanded narrative on tablets smoothly transitions from discussing historical
uses to current functionalities and future possibilities, providing a coherent and
comprehensive overview.
---

Your task is to apply this methodology to the following text and return only the final expanded
version:

<<{input_text}>>
```

Areas of application

- Storage - Persisting “semantic” text content that maintains the original text idea
- Reduce embedding search false positives when applying traditional RAG (*hypothesis, need to check*)
- Improve accuracy of needle in a haystack operations (*hypothesis also need to be checked*)
    - https://blog.langchain.dev/multi-needle-in-a-haystack/
    - https://github.com/gkamradt/LLMTest_NeedleInAHaystack

- …?