CUNY GRADUATE CENTER | QUEENS COLLEGE

AACL-IJCNLP 2026 · HENGQIN, CHINA

Text-Preserving Lossy Text Compression

A Study of Strategic Deletion and LLM Reconstruction

Yuchun Zou[1] · **Junhong Tong**[2] · **Jun Li**[1,2]
[1]CUNY Graduate Center [2]CUNY Queens College

Paper · code · data
github.com/junli-cuny/lossy-text-compression-benchmark

IN ONE SENTENCE

Lossless compression keeps every byte but saves little on natural language, so we delete parts of the text and let an LLM rebuild the original from what remains.

How it works

An encoder deletes characters or words until only a fraction r_{keep} of the text remains. An LLM decoder then rebuilds the original from what is left, the skeleton. We use an LLM because many texts shrink to the same skeleton, so decoding means guessing the most likely original, which is what a language model is trained to do. The decoder runs only on the few chunks that are read back. Below is the opening of a verbatim BBC News article at r_{keep} = 0.5.

1 Source text

~~Film~~ stars ~~from~~ across the ~~globe are preparing~~ to walk the ~~red~~ carpet ~~at~~ this year's Bafta ~~award ceremony~~...The ~~2005 Orange~~ British ~~Academy Film~~ Awards ~~are~~ being ~~held~~ at The ~~Odeon in~~ London's Leicester Square~~,~~ ~~A host~~ of Hollywood stars~~,~~ including Cate Blanchett, Leonardo DiCaprio, ~~Keanu~~ Reeves and Richard ~~Gere~~, ~~are expected to~~ attend ~~Saturday's~~ ceremony~~.~~

The encoder deletes words, spaces, and punctuation by how common they are. Struck-out text is deleted.

encode →

2 Skeleton: what gets stored

```
starsacrossthe to walk the  carpet this
year'sBafta...The  BritishAwards  beingatThe
Londons Leicester Square  of Hollywoodstars
includingCateBlanchett,Leonardo DiCaprio,
Reeves andRichard ,    attendsceremony
```

About half of the article remains. It is still plain text, so it can be searched, edited, and zipped.

decode →

3 LLM reconstruction

Stars from across the globe are set to walk the red carpet at this year's BAFTAs. The EE British Academy Awards are being held at The Odeon in London's Leicester Square, with a host of Hollywood stars, including Cate Blanchett, Leonardo DiCaprio, Keanu Reeves, and Richard Gere, expected to attend the ceremony.

An LLM sees only the skeleton. Green words were deleted and came back exactly. One slip remains, because "2005 Orange" became "EE", a name the article never uses.

ENCODERS

Level 1 · characters, no model.

Step deletion. It removes every k-th character. Three random variants draw the positions instead.

Level 2 · words, no model.

WordLen. It edits words by shape. It drops vowels and short words, then trims words to stems.

WordFreq-3 and WordFreq-6. WordFreq-3 sorts words into three Zipf frequency classes and splits the budget across them in proportion. It drops whole words at random within each class. WordFreq-6 adds buckets for punctuation, other tokens, and whitespace.

LP. A linear program spreads deletions over these six buckets to maximize estimated BERTScore.

Level 3 · runs GPT-2 at the encoder.

Entropy. It drops the most predictable tokens first.

Entropy-LP and Hybrid-α. They combine surprisal with the LP allocation or with frequency rank.

Comparisons.

LLMLingua-2. A trained token classifier from prompt compression, tested as a drop-in encoder.

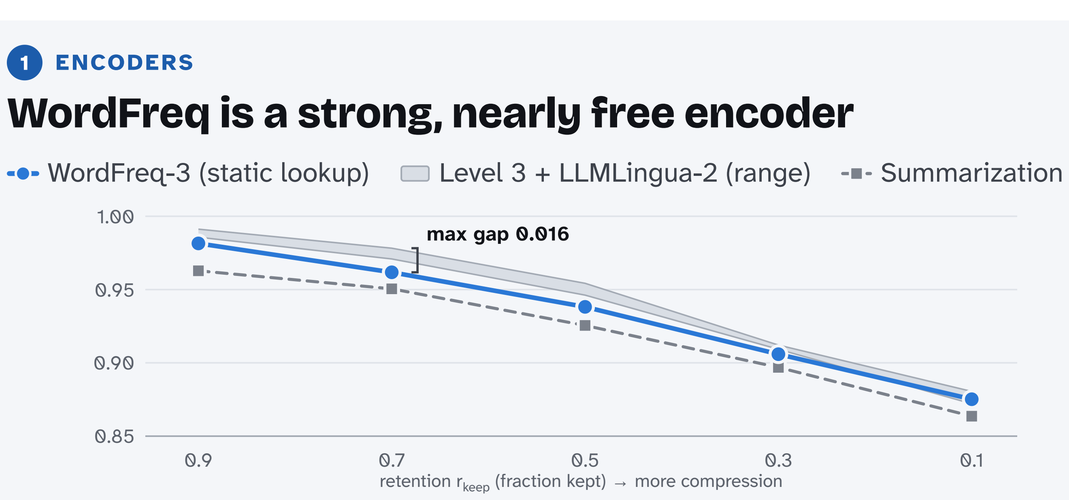
Summarization. It shortens the text to the same length and leaves no skeleton.

DECODERS Google Gemini Flash decodes zero-shot. We fine-tune Llama-3.2-3B with QLoRA, and Qwen2.5-3B, so decoding can run locally.

EVALUATION We test 9 rates from 0.9 down to 0.1 and score with BERTScore, ROUGE-L, and entity recall. People judge facts blind, and two LLM raters scale up their judgments. BBC News is the main corpus, with WikiText, Reddit, and three Chinese corpora.

1 ENCODERS

WordFreq is a strong, nearly free encoder



The gap is largest at mild-to-moderate compression. By r_{keep} = 0.3 it is at most 0.006, and at 0.1 WordFreq falls inside the band.

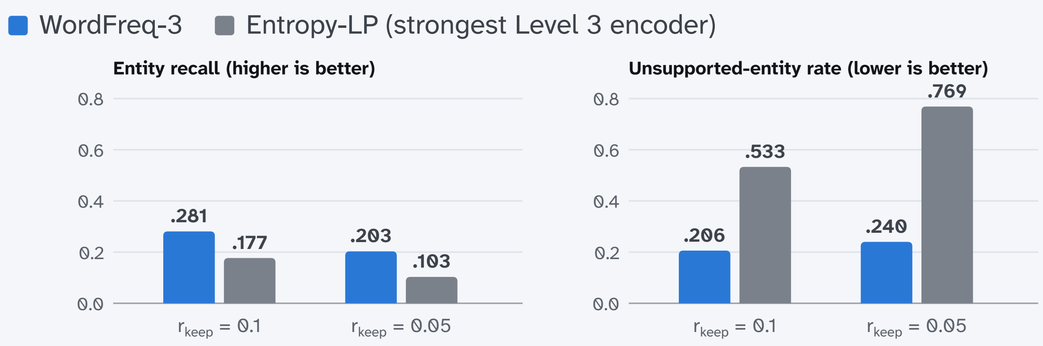
~1 ms
WordFreq, per 512-char chunk

~15 ms
GPT-2 surprisal, GPU

~130 ms
GPT-2 surprisal, CPU

2 ENTITIES

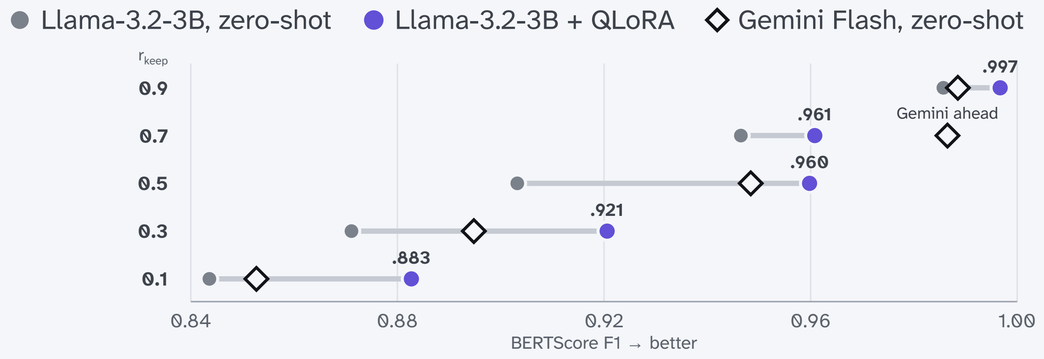
At low retention, WordFreq keeps more of the named entities



Level 3 encoders gain most at mild-to-moderate compression. At r_{keep} = 0.05, WordFreq keeps about twice as many of the original entities as the best of them, at a third of the unsupported-entity rate.

3 DECODERS

A fine-tuned 3B decoder rivals Gemini



QLoRA fine-tuning of Llama-3.2-3B beats zero-shot Gemini Flash at 4 of 5 rates. Gemini leads only at r_{keep} = 0.7. With the skeleton fixed, adaptation adds 0.011–0.056 BERTScore, so reconstruction can run fully on a local machine.

4 EVALUATION

BERTScore overstates what survives

We ran a blind human study on WordFreq-6 and LP reconstructions at r_{keep} = 0.1.

≈0.84
BERTScore F1: reads as close to the source

0.14
ROUGE-L F1: most of the wording is gone

3%
of the facts preserved, judged by people

Similarity still tracks facts overall (Spearman ρ = 0.776 over 54 items). But it stays high at low retention, where fluent text replaces lost content. Only 4 of 54 items invented a claim outright. Most losses are omissions and distortions.

When you evaluate reconstruction, also report word overlap (such as ROUGE-L) and fact preservation (such as entity recall).

Choosing an operating point

REGIME	R_{keep}	STORED*	ENTITY RECALL	UNSUPPORTED	SUITABLE FOR
Mild	0.9–0.8	0.43–0.50	0.89–0.95	0.03–0.06	archival, agent memory
Moderate	0.7–0.5	0.29–0.39	0.65–0.88	0.07–0.15	gist: logs, previews
Aggressive	0.3–0.1	0.08–0.21	0.19–0.54	0.16–0.26	a stress test, not an operating point

*Skeleton after zlib, as a fraction of the original bytes (zlib on the original alone: 0.513). Bands span WordFreq and LP on BBC News.

In practice

Start with WordFreq. It needs no encoder model, and it is the most robust choice once $r_{keep} \le 0.3$.

Switch near $r_{keep} \approx 0.5$. There, LP or a surprisal hybrid can pay off if a model can run at the encoder.

Re-check on your domain. The best rule shifts on WikiText, Reddit, and Chinese corpora, so report word overlap and fact preservation there too.

Contact: yzou@gc.cuny.edu · junhong.tong16@stu-mail.qc.cuny.edu · jun.li@qc.cuny.edu