

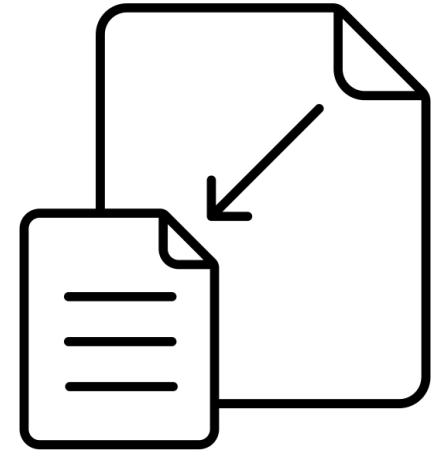
Semantic Compression: Streamlining Text Transmission with Large Language Models

Luning Yang, Yacun Wang
Group: 10

December 2024

Text Compression

- Massive Data: Efficient Transmission, Reduced Storage
- Examples
 - Cached Knowledge Bases
 - IoT Devices
 - Limited LLM context window (Gilbert et al., 2023, Fei et al., 2024)
 - ...



Created by Funtasticon
from Noun Project

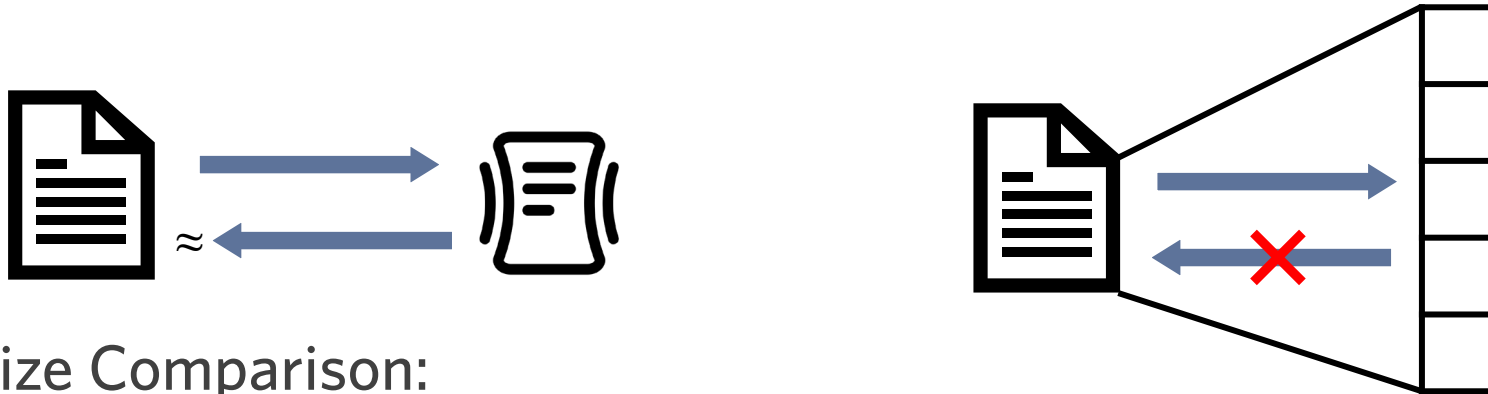
Traditional Algorithms (Lossless)

NYT	Algorithm	Compression Ratio	Information Retained
gzip	DEFLATE (LZ77 + Huffman)	52.08%	100%
bzip2	RLE, BWT, ...	52.24%	100%

- Algorithm Base: String Pattern
- Compression Ratio (CR) = $1 - \text{Compressed Size} / \text{Original Size}$




Compression vs. Embedding

- Compression: Minimize Size, Preserve Integrity
- Embedding: One-Way Latent Mapping



- Size Comparison:
 - Original: 2 bytes (avg) x 2000 characters \approx 4KB
 - Embedding: 8 bytes (float64) x 500 dimensions \approx 4KB
 - Lossless Compression: 2 bytes (avg) x ~1000 characters \approx 2KB

Motivation

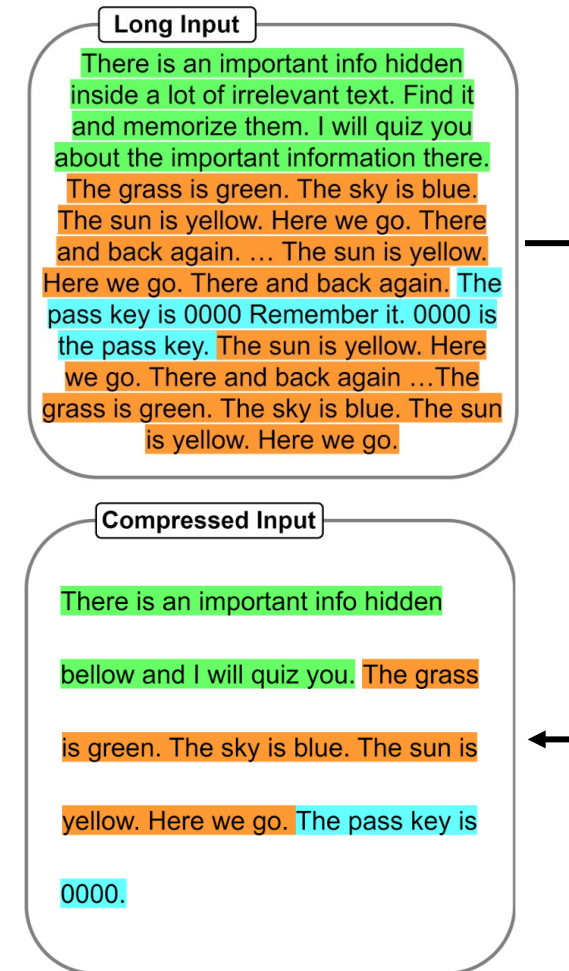
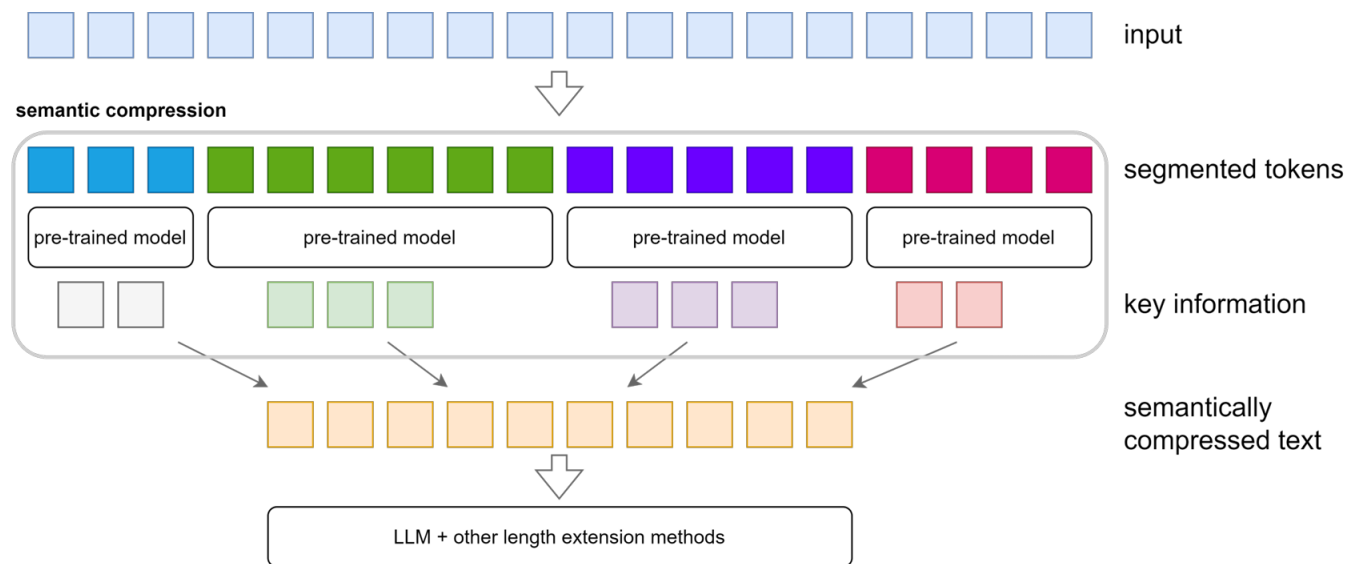
- Lossless vs. Lossy: **Lossy** sometimes acceptable
→ Semantic Compression
- Examples
 - Cached Knowledge Bases 
 - IoT Devices 
 - Limited LLM context window 
- **LLM** Capability: Natural Language Understanding

Related Literature

- Gilbert et al., 2023:
 - Semantic Compression using LLMs on short stories (~76%) and code generation
 - Aim to fit in LLM input token limit

Related Literature

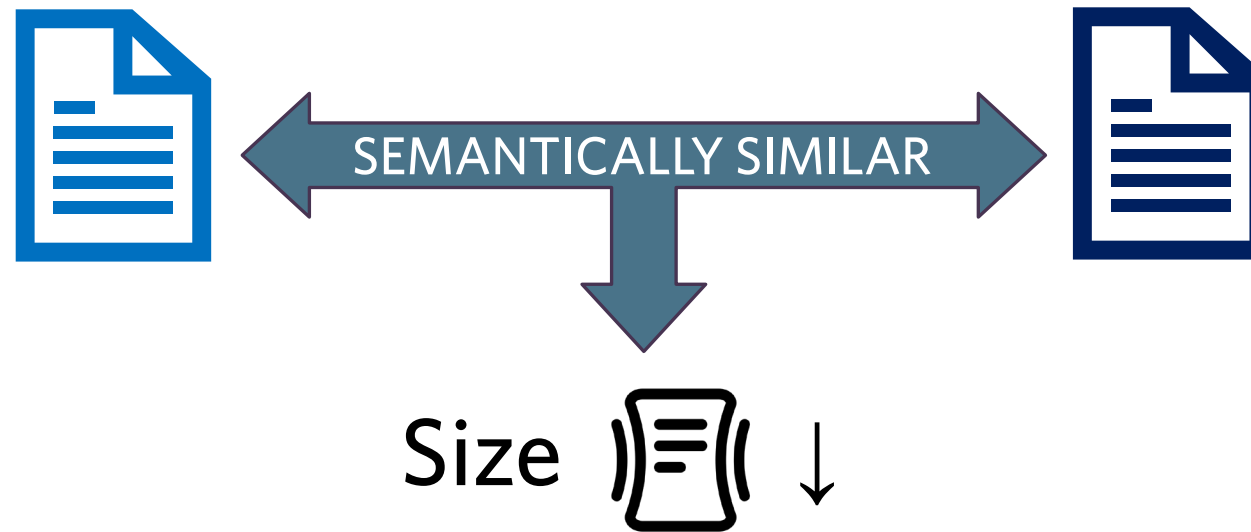
- Gilbert et al., 2023
- Fei et al., 2024
 - Extend LLM context window
 - Improved performance on end tasks



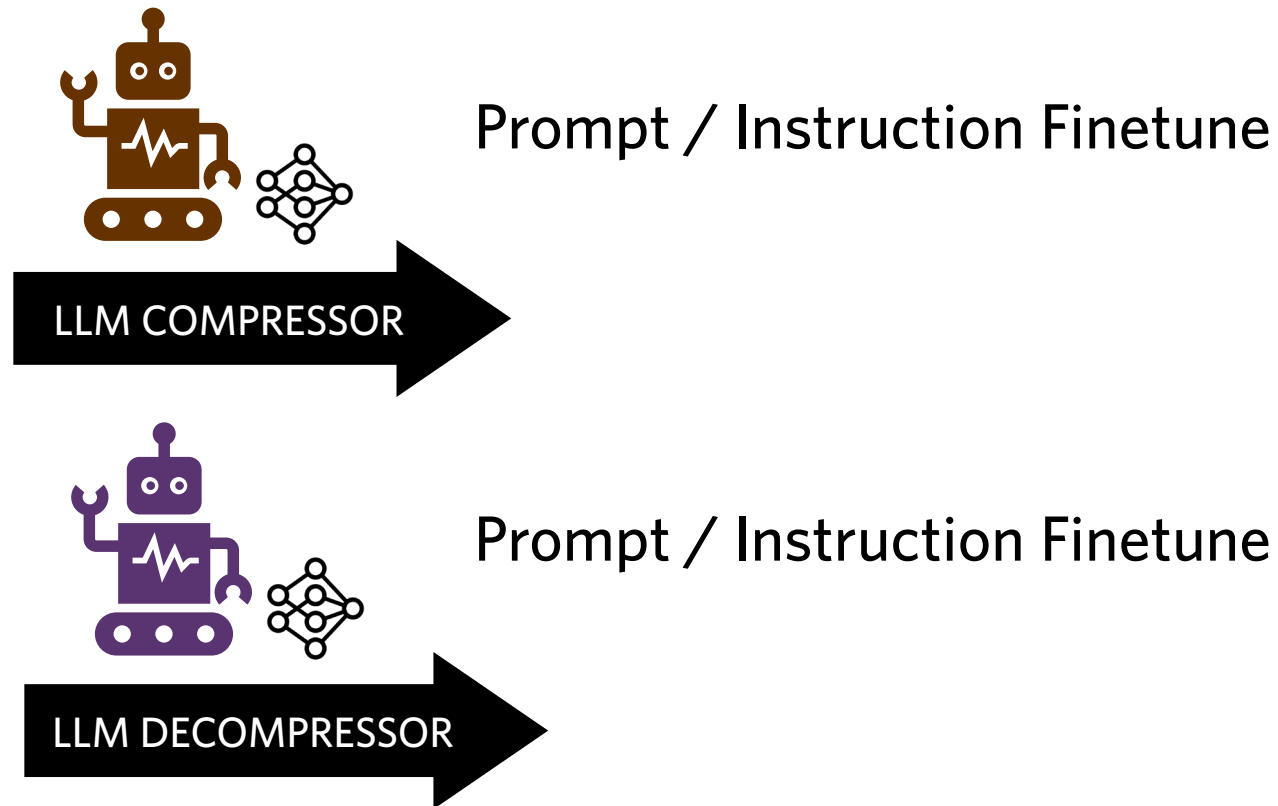
Approach



Objectives



Baseline Approach



Baseline Results (GPT-4o-mini)

- Decompressed Similarity: SBERT Embeddings + Cosine Similarity

NYT	Compression Ratio	SBERT Embeddings
gzip/bzip2	Around 52%	1.000
Prompting	67.7%	0.777

Baseline Results (GPT-4o)

- Decompressed Similarity: GPT Embeddings + Cosine Similarity

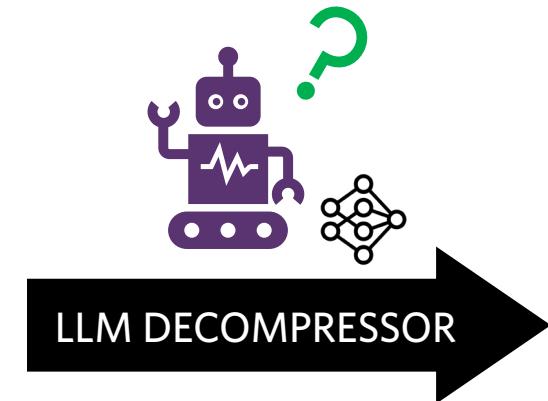
NYT	Compression Ratio	Decompressed Similarity
gzip/bzip2	Around 52%	1.000
Prompting	67.7%	0.881
Finetuned Decompressor		0.940

Baseline Limitations

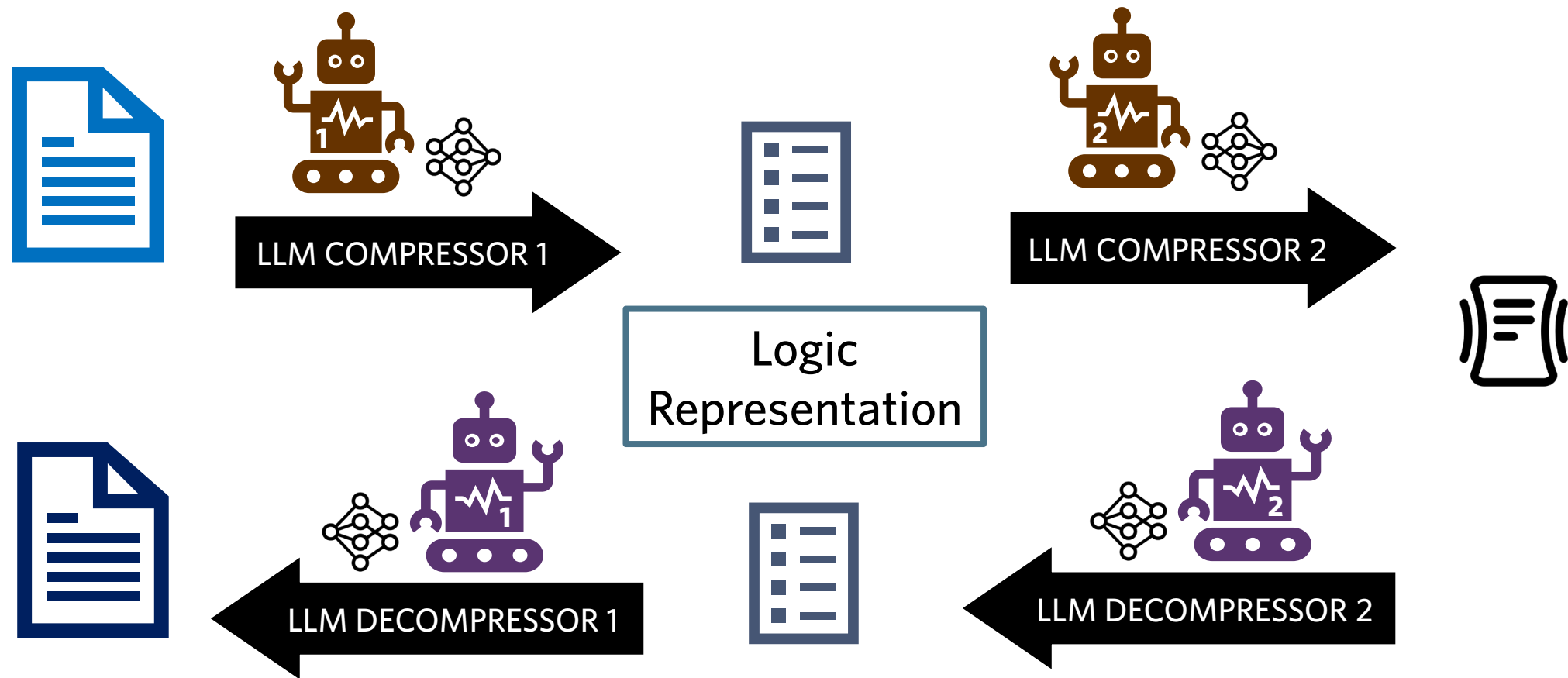
- Lack of guidance
- Decompression Failure Rate: **35.8%**

Example Compressed Text:

st.l—pk🏏cz🐢chz📦⇒2👁️🦅🏏🏏1🌙s🏏💡g🍺p↑ΠOXY1⚡-⌚6-
1⌚🏠⇒🐢8📺⇒📦3oP🎯🛩️13❌🔄6👁️📶4❌NG🔄⌚❌📈
⌚👉📶n🌿💥d❤️50👉👉🍂31📺🖋️?🍌👉🖋️3(2.72)*7G🐢3(R
z)~5🏑🏏🌟📶🔄100❌🏆📄🦊👉🐿️👍

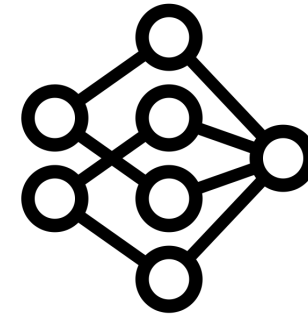


Multi-Stage Compression



Design Choices

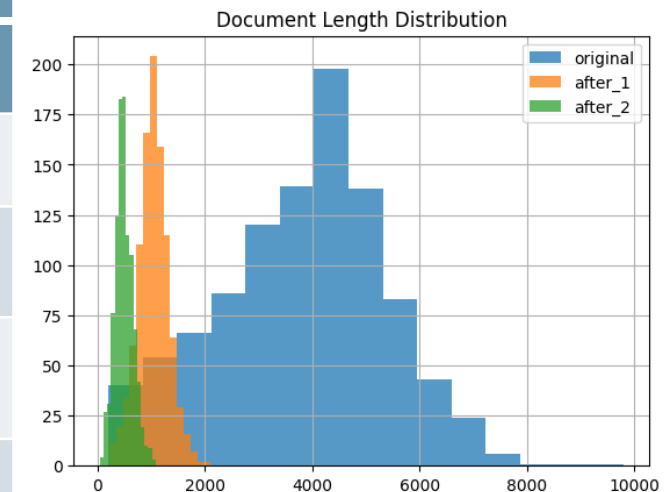
- Intermediate Format: Note-like bullet points
- Prompt Hint: Include Text Genre



Multi-Stage Results (GPT-4o-mini)

- Computational Limitation: Randomly Sampled 1000 documents

NYT	Compression Rate	Decompressed Cosine Similarity	
		GPT Embeddings	SBERT Embeddings
gzip/bzip2	~52%	1.000	1.000
Prompt Baseline	67.7%	-	0.777
Prompt Formatted (Intermediate Stage, After 1)	56.5%	0.754	0.791
Prompt Short (Fully Compressed, After 2)	85.3%	0.783	0.826



Multi-Stage Results (GPT-4o-mini, Downstream Task)

- Topic Classification Accuracy: Direct Prompting














NYT	Coarse (6-Class)	Fine (26-Class)
Original	0.966	0.895
After Compressor 1	0.963	0.909
After Compressor 2	0.960	0.890

Case Study

Original Text:

despite earlier fears of a rainstorm, the weather was clear and breezy, if muggy, on wednesday night for the new york philharmonic 's concert in prospect park, the "most beautiful park in these united states," in the effusive words of marty markowitz, the brooklyn borough president, who addressed the estimated crowd of ,000 people. alan gilbert, the philharmonic's music director, ...
(Length: 4494)

Compressed Text:

NY Phil Concert, Wed Night, Prospect Park, ~10K attend, Alan Gilbert cond, Carter Brey solo, Dvořák Cello Concerto , Tchaikovsky 5th , 14M park concerts, tours , past: Bocelli, Mahler , Gilbert's e-mail , outdoor sound , upcoming: indoor brass @ Staten Island, charity for Sandy Mariah , , Dvořák: intimate , Tchaikovsky: fresh , fireworks , Gilbert: all summer  
(Length: 374)

Decompressed Text:

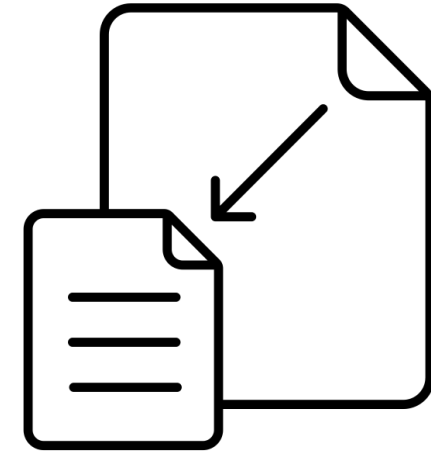
On a beautiful Wednesday night, the New York Philharmonic brought music to life in Prospect Park, attracting an impressive crowd of approximately 10,000 people. Under the direction of acclaimed conductor Alan Gilbert, the summer concert series continued to delight audiences with a captivating program that featured renowned cellist Carter Brey performing Dvořák's beloved Cello Concerto, alongside Tchaikovsky's powerful Symphony No. 5... (Length: 2270)

CR: 91.67%

SBERT Similarity:
0.824

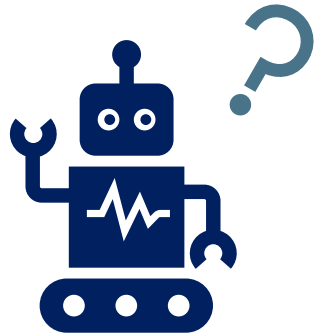
Future Work

- Effects of Model Scale
 - Capabilities of Smaller Model Compression
 - Reduced Computational Cost
- Effects of Intermediate Format
- Effects of Text Genre
- Challenges
 - Controlling LLM output format/length
 - New Data Less Likely to be Pretrained On
 - Style Lost



Created by Funtasticon
from Noun Project

Q & A



References

- Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2024. Extending context window of large language models via semantic compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5169–5181, Bangkok, Thailand. Association for Computational Linguistics.
- Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.
- Henry Gilbert, Michael Sandborn, Douglas C. Schmidt, Jesse Spencer-Smith, and Jules White. 2023. Semantic compression with large language models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8.

Credit

- Page 2: compression by Funtasticon from `Noun Project` (CC BY 3.0)
- Compression by Funtasticon from `Noun Project` (CC BY 3.0)
- Neural Network by Cécile Lanza Parker from `Noun Project` (CC BY 3.0)