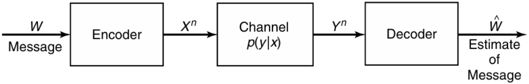
LLM-Based Codes for Deletion Channels

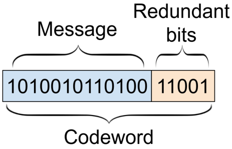
Rohit Bhagat   Salim El Rouayheb

Rutgers University

July 17, 2025

Review of Deletion Channels





$\mathcal{X}$ = channel input alphabet $\quad \mathcal{Y}$ = channel output alphabet

Deletion Channel: Every symbol X_i in message $X \in \mathcal{X}^n$ is dropped from the message with probability p

Example: $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$

$$X = 1010$$
$$Y = 110$$

Shannon's Noisy-Channel Coding Theorem

Every channel has a capacity $C = \sup_{p(X)} I(X; Y)$.

Theorem (Shannon's Noisy-Channel Coding Theorem)

For any rate $R < C$, there exists a code that achieves R with arbitrarily small probability of error.

Notes:

- Shannon did not tell us *how* to achieve this rate
- The capacity of the deletion channel is not known

Project Vision

Use the English alphabet as input/output alphabet

$\mathcal{X} = \{$A, B, C, ..., X, Y, Z, a, b, c, ..., x, y, z$\}$
$\mathcal{Y} = \{$A, B, C, ..., X, Y, Z, a, b, c, ..., x, y, z$\}$

Exploit inherent redundancy:
"DIACS REUis th coolesresearh pogam ver ceatd!"

Use LLMs as a tool to recover text with deletions

Is GPT Better Than You?

Deleted text ($p = 0.30$):

“heele wee on of three easvewed as bein inthe nnior Rdgers.”

Is GPT Better Than You?

Deleted text ($p = 0.30$):
"heele wee on of three easvewed as bein inthe nnior Rdgers."

Original:
"The Steelers were one of three teams viewed as being in the running for Rodgers."

Is GPT Better Than You?

Deleted text ($p = 0.30$):
“heele wee on of three easvewed as bein inthe nnior Rdgers.”

Original:
“The Steelers were one of three teams viewed as being in the running for Rodgers.”

GPT-recovered:
“The Steelers were one of three teams viewed as being in the running for Rodgers.”

Creating Datasets

Need for large amounts of plaintext

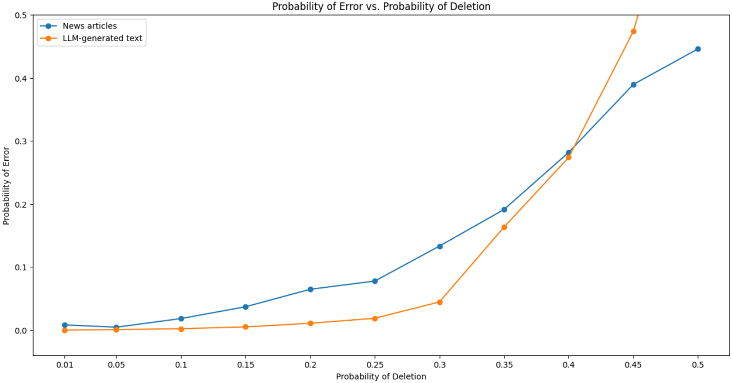
Method 1: Scraping recent news articles

- "After Saturday's 3-1 loss to Turkey..."
- "An international team led by Dr. Andy Rivkin..."
- "<b>What seafood should you avoid?</b>"

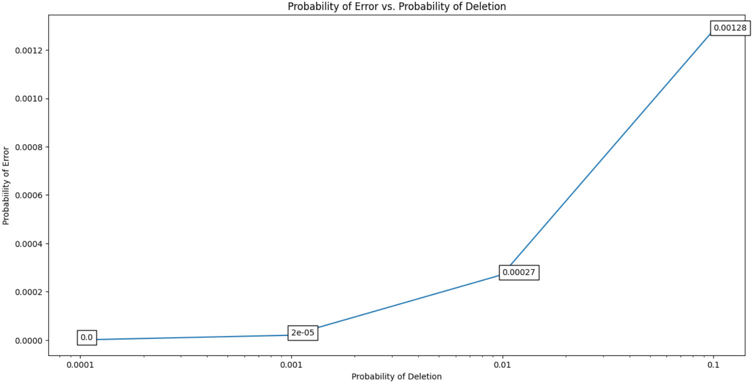
Method 2: Generating "clean" texts using LLMs

- "...divide cashflows into tranches..."
- "...crucial for companies operating loally..."

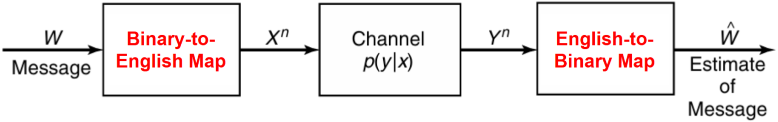
GPT Performance on Datasets



GPT Performance on Datasets



Mapping Binary to a Language



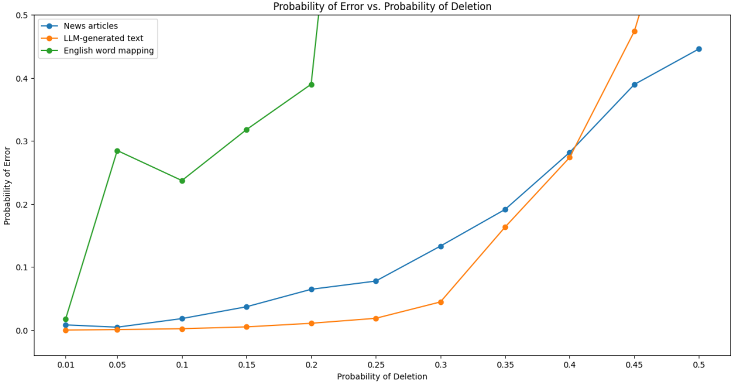
Idea 1: Map all 2^b binary strings of length b to English *words*

- 0000 → apple
- 0001 → banana
- ...

Idea 2: Map all 2^b binary strings of length b to English *sentences*

- 0000 → The sky is blue.
- 0001 → P equals NP.
- ...

GPT Performance Using Mapping



Next Steps

- Explore more efficient mappings from binary to some structured language
- Explore alternatives to edit distance (semantic distance)
- Automatic identification of parts of text that cannot be inferred from context
- Theoretical guarantees on performance

Acknowledgments

I would like to thank

- My mentor, Dr. Salim El Rouayheb, for advising this project.
- The DIMACS REU 2025 program, in which this research is being conducted.
- NSF grants CCF-2447342 and CNS-2148182 for supporting this research.

References

T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*.
Wiley-Interscience, July 2006.

M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, no. none, pp. 1 – 33, 2009.