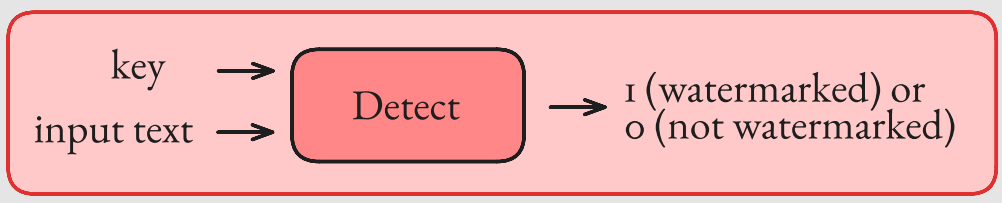
LLM Watermarking

Antonín Jarolím, Vojtěch Eichler
xjarol06@fit.vutbr.cz, xeichl01@fit.vutbr.cz

Watermark and Detection Process

- **Watermark** ($\mathcal{M}$): procedure that outputs Watermarked model $\hat{\mathcal{M}}$, and detection key k

- **Detect** (k, **y**): takes input detection key k and sequence **y**, then outputs 1 (indicating it was AI-generated) or 0 (indicating it was human-generated)



GumbelSoft watermark process

Input: prompt r, LLM $\mathcal{M}$, temperature τ
Output: Watermarked sequence $w_1, \ldots, w_T$

1: **for** $t = 1, \ldots, T$ **do**
2: Logits $l_t \leftarrow \mathcal{M}(r, w_{1,\ldots,t-1})$
3: Watermark key $\xi_t \leftarrow$ hash context to a Gumbel-distributed vector
4: $w_t \leftarrow$ sample from softmax$((\xi_t + l_t)/\tau)$
5: **end for**
6: **return** $[w_1, \ldots, w_T]$

Red-green watermark process

- boost logits of randomly (based on key) generated vocabulary (green) split [Kirchenbauer et al., 2024]

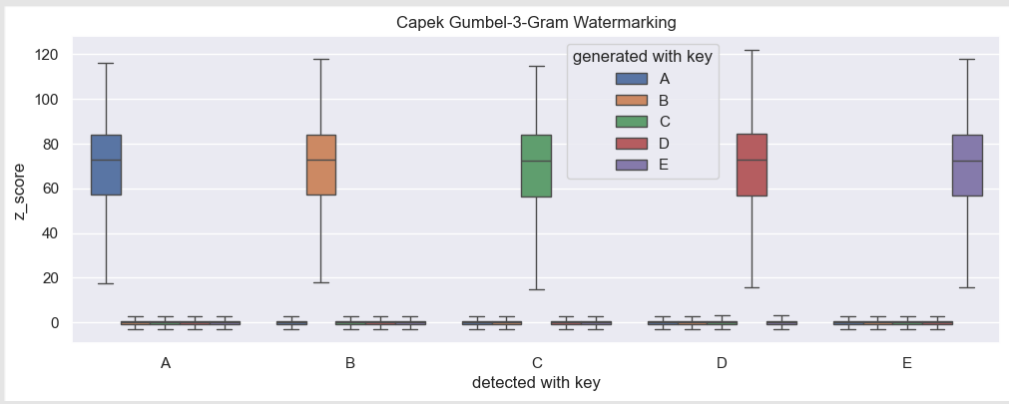
…V roce 1989 byla nalezena v anglickém městě Cambridge . Během zkoumání se zjistilo , že čelist patřila muži , který zemřel o 300 let dříve , než se začalo s ko páním do země . V roce 189 2 prováděl v Cambridge výzkum geo log Charles Wil kins , na základě jeho zkoumání byl vědec považován za původního majitele . …

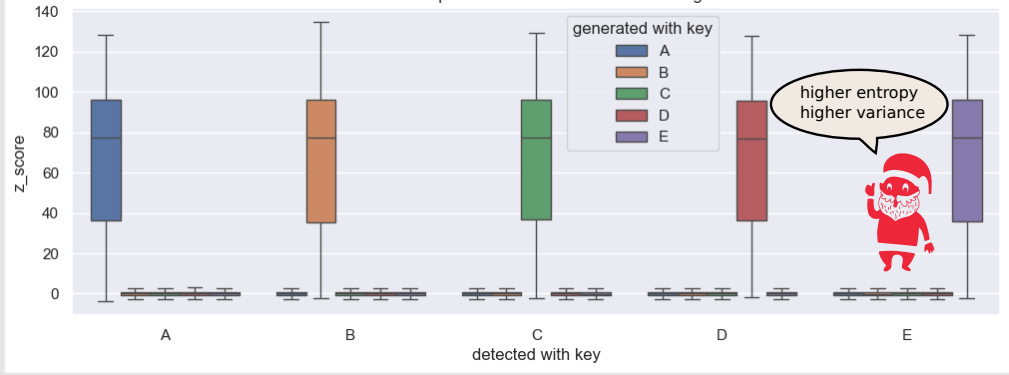
…V roce 1989 byla Ste y re gg ská tvrz zbo řena a vodní zdroj obnoven . Voda pro tvrz př ité ká z nedalekého K ast en re it ského potoka přes upravenou zemní studá nku . Původní tvrz stála přímo u bro du přes K ast en re it ský potok v blízkosti Ste y re gg ského mostu . Po přestavbě tvrze po povo dni v 16 . …

Datasets

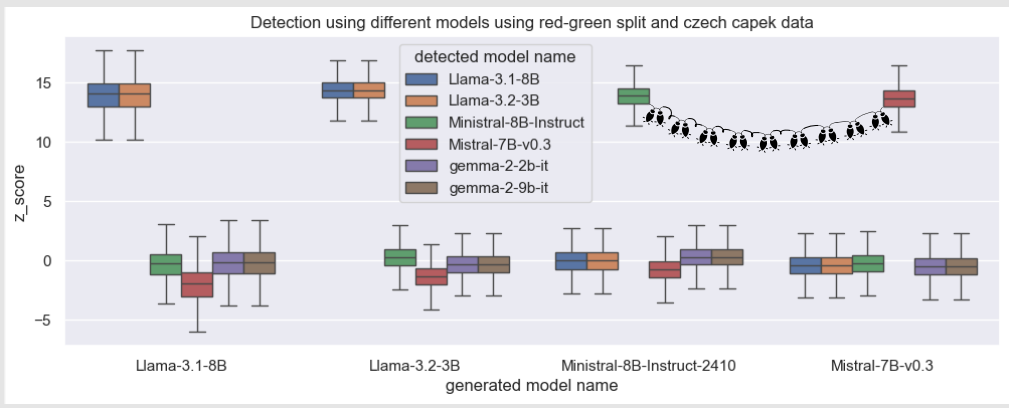
- **OpenGen [Krishna et al., 2023] and Capek (Czech books) [Čermák et al., 2007]**: dialogue generation and story telling creative completion → higher entropy
- **SQuAD [Rajpurkar et al., 2016]**: fact-based QA format narrows LLM responses → lower entropy

Task 1: Detection with different keys

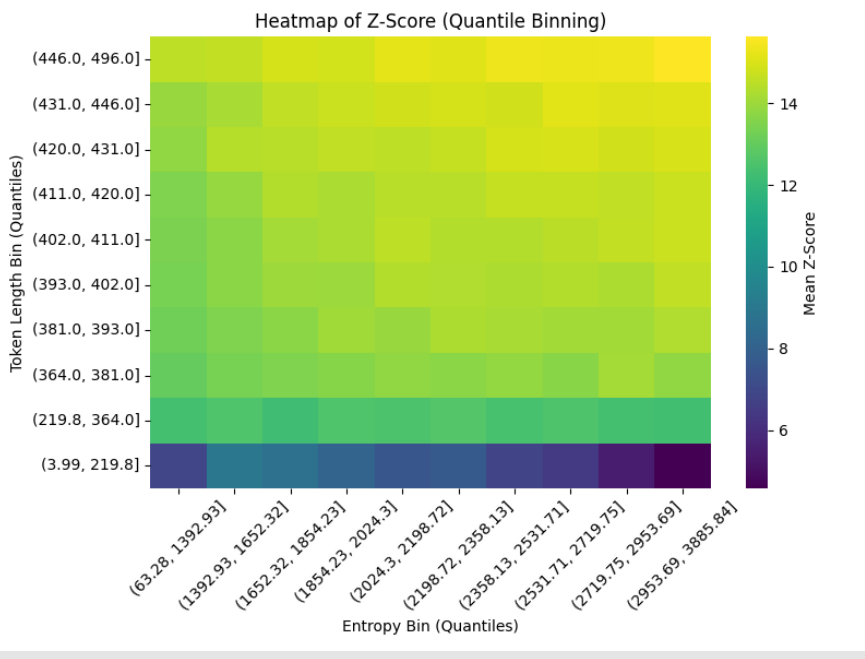




Task 2: Detection with different models



Task 3: Exploring effects of length and entropy on detectability



References

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2024). A watermark for large language models.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Čermák, F. et al. (2007). Capek: Korpus pouze vlastních textů karla Čapka. Ústav Českého národního korpusu FF UK, Praha.
Released corpus.