LLM Seminar Topics WiSe 2023/24

Kathy Hämmerl
haemmerl@cis.lmu.de

CIS, LMU

28 November 2023

Table of Contents

Table of Contents

Decoding Strategies

- A language model outputs likelihood predictions, but to get text output we must use a *decoding strategy*.

Decoding Strategies

- A language model outputs likelihood predictions, but to get text output we must use a *decoding strategy*.
- Decoding strongly influences diversity, quality, information content, perplexity, etc. of the output.

Decoding Strategies

- A language model outputs likelihood predictions, but to get text output we must use a *decoding strategy*.
- Decoding strongly influences diversity, quality, information content, perplexity, etc. of the output.
- Discuss how different decoding methods trade off between these aspects, particularly quality and probability.

Decoding Strategies

- A language model outputs likelihood predictions, but to get text output we must use a *decoding strategy*.
- Decoding strongly influences diversity, quality, information content, perplexity, etc. of the output.
- Discuss how different decoding methods trade off between these aspects, particularly quality and probability.
- Why might some strategies be more effective in certain contexts than others?

Decoding Strategies

- A language model outputs likelihood predictions, but to get text output we must use a *decoding strategy*.
- Decoding strongly influences diversity, quality, information content, perplexity, etc. of the output.
- Discuss how different decoding methods trade off between these aspects, particularly quality and probability.
- Why might some strategies be more effective in certain contexts than others?

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On Decoding Strategies for Neural Text Generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Table of Contents

Inequality Between Languages

- Most commercial LLMs are trained on mostly English but have multilingual capabilities

Inequality Between Languages

- Most commercial LLMs are trained on mostly English but have multilingual capabilities
- Smaller data for “other” languages leads to smaller tokenisation, i.e., more pieces per word

Inequality Between Languages

- Most commercial LLMs are trained on mostly English but have multilingual capabilities
- Smaller data for “other” languages leads to smaller tokenisation, i.e., more pieces per word
- Discuss how tokenisation interacts with performance and pricing of commercial APIs. What solutions might be available for providers or users?

Inequality Between Languages

- Most commercial LLMs are trained on mostly English but have multilingual capabilities
- Smaller data for "other" languages leads to smaller tokenisation, i.e., more pieces per word
- Discuss how tokenisation interacts with performance and pricing of commercial APIs. What solutions might be available for providers or users?

Ahia, Orevaoghene et al. 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. CoRR, ArXiv abs/2305.13707

Table of Contents

Factuality in LLMs

- LLMs do not inherently prioritise factual outputs or citations, so fact-checking is very often necessary

Factuality in LLMs

- LLMs do not inherently prioritise factual outputs or citations, so fact-checking is very often necessary
- Attempts to improve this include drawing on knowledge bases, forcing the model to "cite" sources, and more

Factuality in LLMs

- LLMs do not inherently prioritise factual outputs or citations, so fact-checking is very often necessary
- Attempts to improve this include drawing on knowledge bases, forcing the model to "cite" sources, and more
- Discuss how citations can be generated, how reliable they are, and what next steps might be

Factuality in LLMs

- LLMs do not inherently prioritise factual outputs or citations, so fact-checking is very often necessary
- Attempts to improve this include drawing on knowledge bases, forcing the model to "cite" sources, and more
- Discuss how citations can be generated, how reliable they are, and what next steps might be

Gao, Tianyu et al. 2023. Enabling Large Language Models to Generate Text with Citations. In *EMNLP 2023*