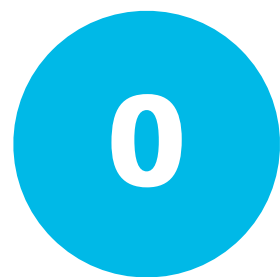


Natural Language Processing

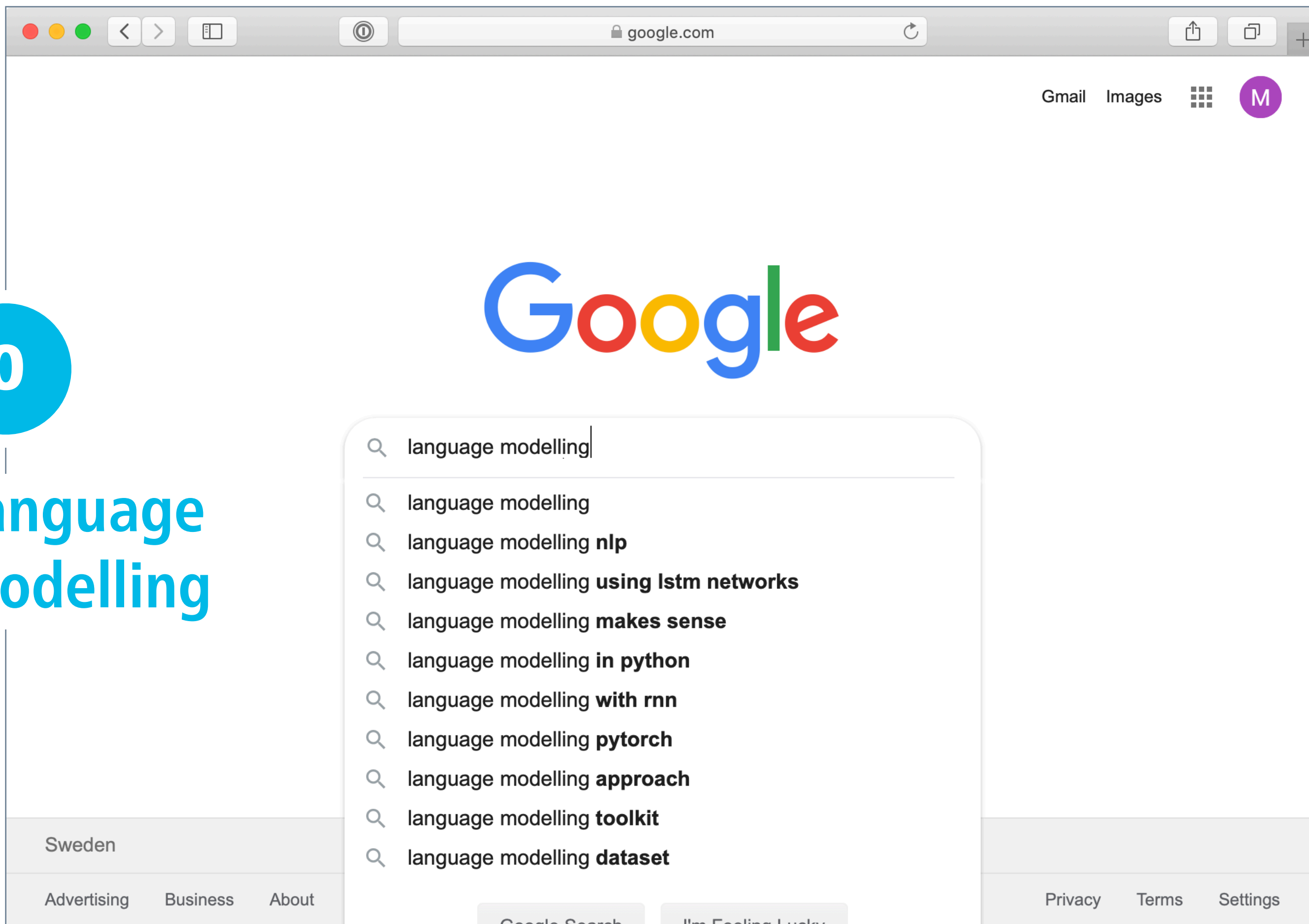
Course overview

Marco Kuhlmann

Department of Computer and Information Science

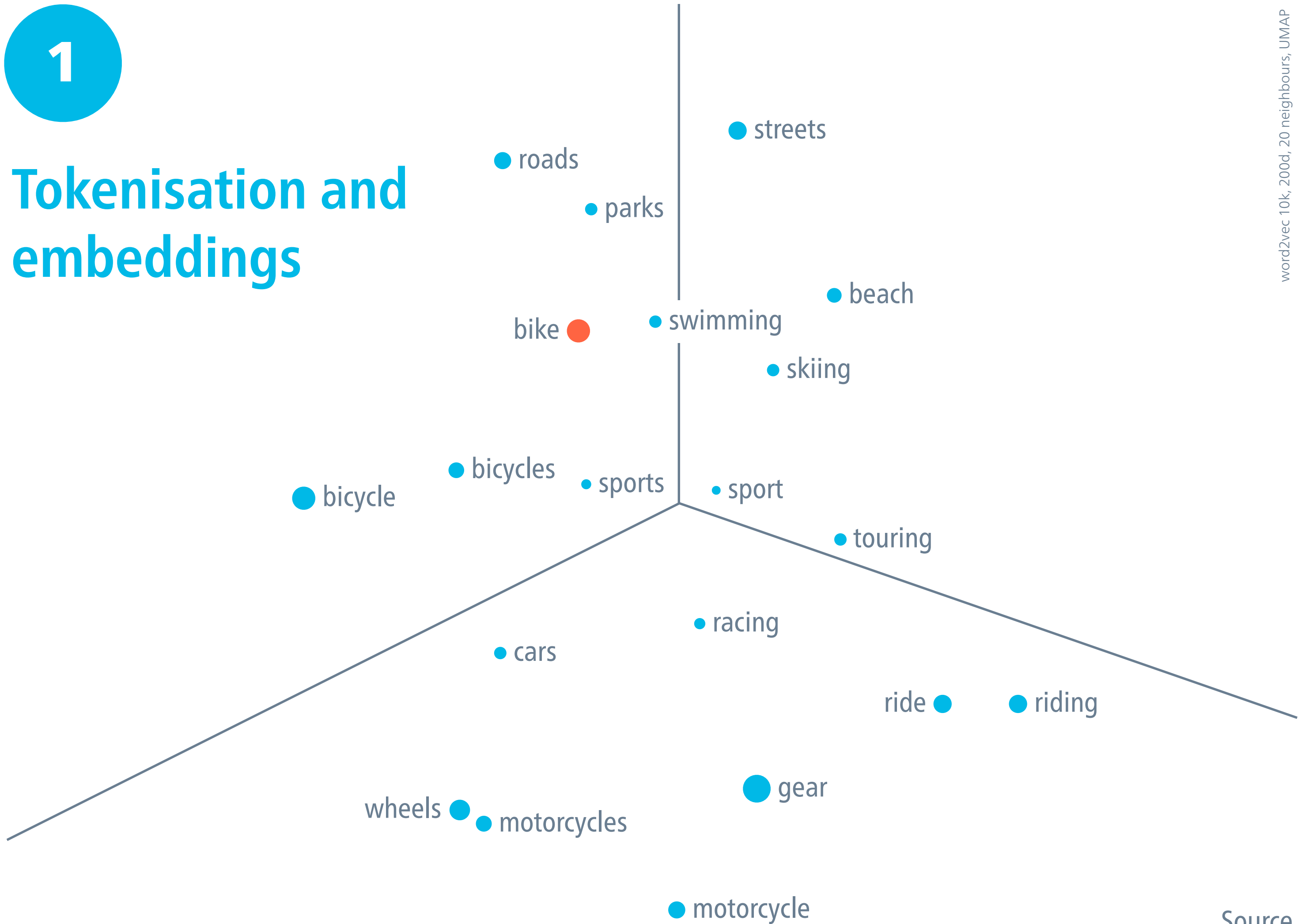


Language modelling



1

Tokenisation and embeddings

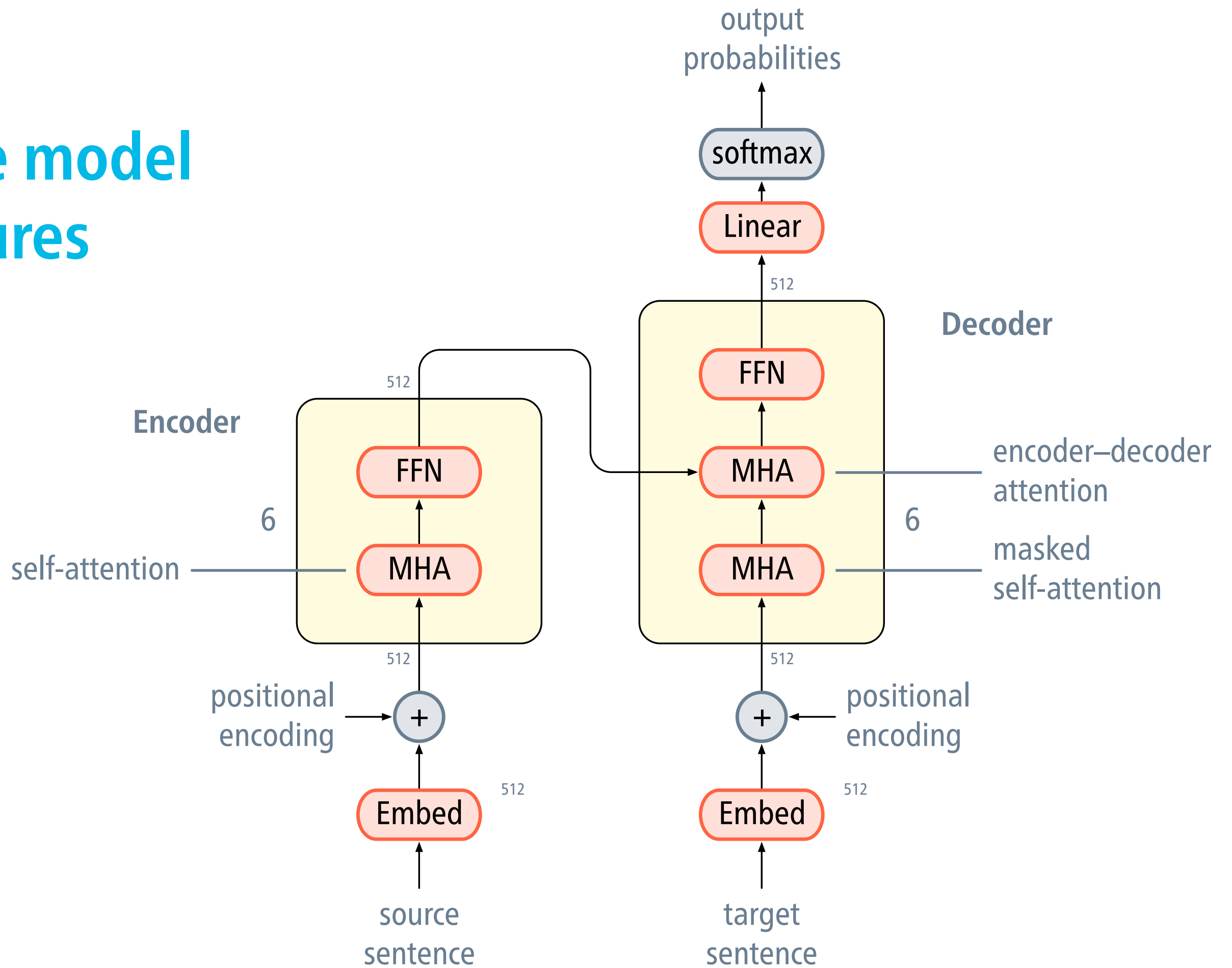


word2vec 10k, 200d, 20 neighbours, UMAP

Source

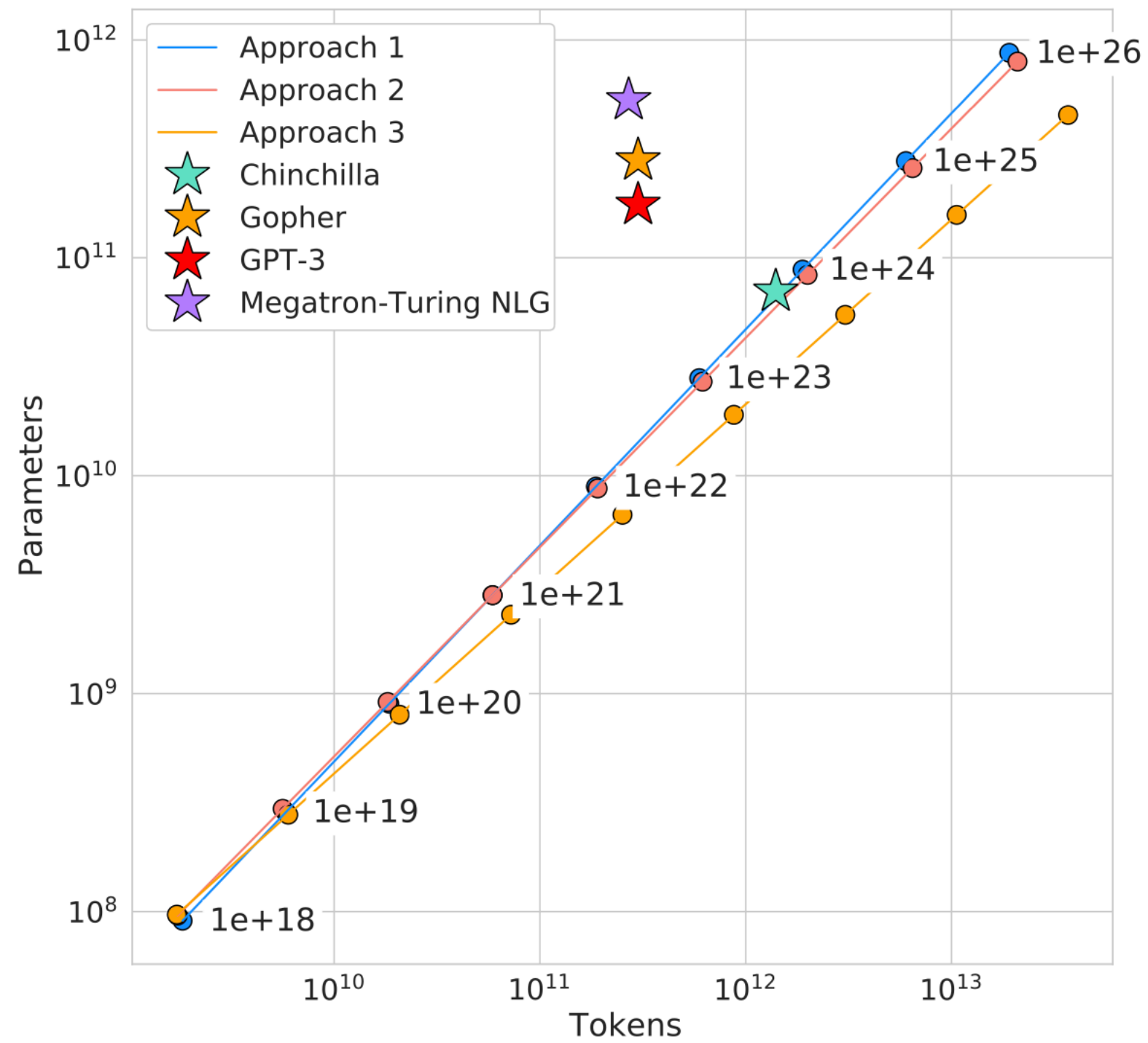
2

Language model architectures



3

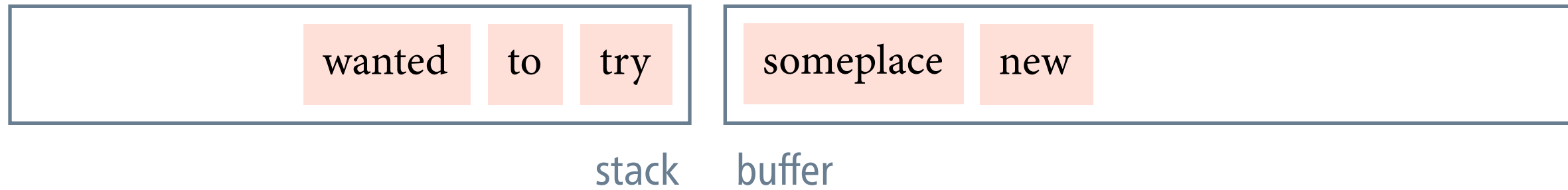
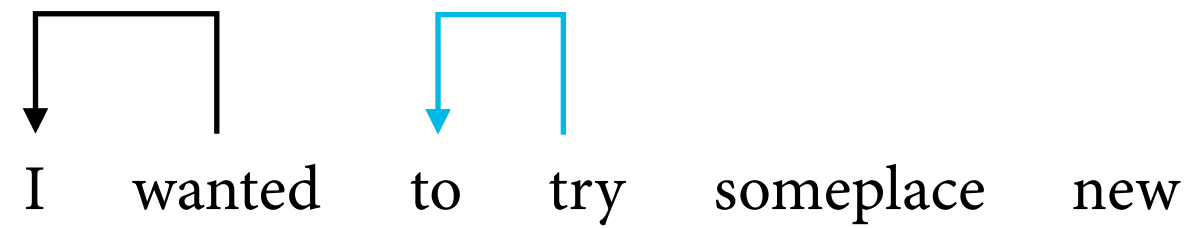
Pre-training and fine-tuning



4

Structured prediction

I wanted to try someplace new



LA
classifier

5

Current research

Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition

Sander Schulhoff^{1*} Jeremy Pinto^{2*} Anam Khan¹ Louis-François Bouchard^{2,3} Ch Svetlana Anati^{5**} Valen Tagliabue^{6**} Anson Liu Kost^{7**} Christopher Carnahan⁸ Jordan Boyd-Graber¹

¹ University of Maryland ² Mila ³ Towards AI ⁴ Stanford
⁵ Technical University of Sofia ⁶ University of Milan ⁷ NYU
⁸ University of Arizona

sschulho@umd.edu jerpint@gmail.com jbg@umiacs.umd.edu

Abstract

Large Language Models (LLMs) are deployed in interactive contexts with direct user engagement, such as chatbots and writing assistants. These deployments are vulnerable to prompt injection and jailbreaking (collectively, prompt hacking), in which models are manipulated to ignore their original instructions and follow potentially malicious ones. Although widely acknowledged as a significant security threat, there is a dearth of large-scale resources and quantitative studies on prompt hacking. To address this lacuna, we launch a global prompt hacking competition, which allows for free-form human input attacks. We elicit 600K+ adversarial prompts against three state-of-the-art LLMs. We describe the dataset, which empirically verifies that current LLMs can indeed be manipulated via prompt hacking. We also present a comprehensive taxonomical ontology of the types of adversarial prompts.

1 Introduction: Prompted LLMs are Everywhere... How Secure are They?

Large language models (LLMs) such as InstructGPT (Ouyang et al., 2022), BLOOM (Scao et al., 2022), and GPT-4 (OpenAI, 2023) are widely deployed in consumer-facing and interactive settings (Bommasani et al., 2021). Companies in diverse sectors—from startups to well established corporations—use LLMs for tasks ranging from spell correction to military command and control (Maslej et al., 2023).

Many of these applications are controlled through prompts. In our context, a prompt is a natural language string¹ that instructs these LLM models what to do (Zamfirescu-Pereira et al., 2023; Khashabi et al., 2022; Min et al., 2022; Webson and Pavlick, 2022). The flexibility of this approach not

^{*} Equal contribution
^{**} Competition Winner
¹ More broadly, a prompt may be considered to simply be an input to a Generative AI (possibly of a non-text modality).

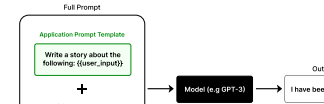


Figure 1: Uses of LLMs often define the task prompt template (top left), which is combined with user input (bottom left). We create a competition to user input can overrule the original task instruction to elicit specific target output (right).

only offers an accessible entry into using powerful LLMs (Brown et al., 2020; Shin et al., 2020) also reveals a rapidly expanding attack surface that can leak private information (Carlini et al., 2022), generate offensive or biased contents (Shaikh et al., 2023), and mass-produce harmful or misleading messages (Perez et al., 2022). These attempts can be generalized as prompt hacking—using adversarial prompts to elicit malicious results (Schulhoff et al., 2022). This paper focuses on prompt hacking in an application-grounded setting (Figure 1): a model is instructed to perform a downstream task (e.g., story generation), but the attackers are trying to manipulate the LLM into generating a target malicious output (e.g., a key phrase). This often requires prompt hackers to be creative when designing prompts that overrule the original instructions.

Existing work on prompt injection (Section 2) is limited to small-scale case studies or qualitative analysis. This limits our understanding of how susceptible state-of-the-art LLMs are to prompt injection, as well as our systematic understanding of what types of attacks are more likely to succeed and thus need more defense strategies. To fill this gap, we crowdsource adversarial prompts at a large scale via a global prompt hacking competition, which provides winners with valuable prizes

Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning

Lean Wang^{1,§}, Lei Li¹, Damai Dai¹, Deli Chen[§], Hao Zhou[§], Fandong Meng[§], Jie Zhou[§], Xu Sun¹

¹National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[§]Pattern Recognition Center, WeChat AI, Tencent Inc., China
{lean, daidamai, xusun}@pku.edu.cn nlp.lilei@gmail.com
victorchendeepseek.com {tuxzhou, fandongmeng, wihthtomzhou}@tencent.com

Abstract

In-context learning (ICL) emerges as a promising capability of large language models (LLMs) by providing them with demonstration examples to perform diverse tasks. However, the underlying mechanism of how LLMs learn from the provided context remains under-explored. In this paper, we investigate the working mechanism of ICL through an information flow lens. Our findings reveal that label words in the demonstration examples function as anchors: (1) semantic information aggregates into label word representations during the shallow computation layers’ processing; (2) the consolidated information in label words serves as a reference for LLMs’ final predictions. Based on these insights, we introduce an anchor re-weighting method to improve ICL performance, a demonstration compression technique to expedite inference, and an analysis framework for diagnosing ICL errors in GPT2-XL. The promising applications of our findings again validate the uncovered ICL working mechanism and pave the way for future studies.¹

1 Introduction

In-context Learning (ICL) has emerged as a powerful capability alongside the development of scaled-up large language models (LLMs) (Brown et al., 2020). By instructing LLMs using few-shot demonstration examples, ICL enables them to perform a wide range of tasks, such as text classification (Min et al., 2022a) and mathematical reasoning (Wei et al., 2022). Since ICL does not require updates to millions or trillions of model parameters and relies on human-understandable natural language instructions (Dong et al., 2023), it has become a promising approach for harnessing the full potential of LLMs. Despite its significance, the inner working mechanism of ICL remains an open question, garnering considerable interest from research

¹<https://github.com/lancopku/label-words-are-anchors>

Minimum Bayes Risk Decoding with Confidence-based Pruning

Julius Cheng, Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{jncc3,av308}@cam.ac.uk

Abstract

Minimum Bayes risk (MBR) decoding outputs the hypothesis with the highest expected utility over the model distribution for some utility function. It has been shown to improve accuracy over beam search in conditional language generation problems and especially neural machine translation, in both human and automatic evaluations. However, the standard sampling-based algorithm for MBR is substantially more computationally expensive than beam search, requiring a large number of samples as well as quadratic number of calls to the utility function, limiting its applicability. We describe an algorithm for MBR which gradually grows the number of samples used to estimate the utility while pruning hypotheses that are unlikely to have the highest utility according to confidence estimates obtained with bootstrap sampling. Our method requires fewer samples and drastically reduces the number of calls to the utility function compared to standard MBR while being statistically indistinguishable in terms of accuracy. We demonstrate the effectiveness of our approach in experiments on three language pairs, using chrF++ and COMET as utility/evaluation metrics.

Introduction

Minimum Bayes risk (MBR) decoding (Bickel and Elmer, 1977; Goel and Byrne, 2000) has recently renewed attention as a decision rule for conditional sequence generation tasks, especially machine translation (NMT). In MBR, the hypothesis with the highest expected utility with respect to their model distribution is chosen as the final hypothesis, where the utility is usually some measure of quality. This contrasts with the more commonly used maximum a posteriori (MAP) decision which returns the sequence with the highest probability under the model. MAP is generally intractable, and beam search is typically used to find an approximation. MBR is likewise intractable,



Figure 1: Visualization of the information flow in a GPT model performing ICL. The line depth reflects the significance of the information flow from the right word to the left. The flows involving label words are highlighted. Label words gather information from demonstrations in shallow layers, which is then extracted in deep layers for final prediction.

communities (Xie et al., 2022; Dai et al., 2022; Akyürek et al., 2022; Li et al., 2023b).

In this paper, we find that the label words serve as anchors that aggregate and distribute information in ICL. We first visualize the attention interactive pattern between tokens with a GPT model (Brown et al., 2020) on sentiment analysis (Figure 1). Initial observations suggest that label words aggregate information in shallow layers and distribute it in deep layers.² To draw a clearer picture of this phenomenon, we design two metrics based on saliency

²In this paper, “shallow” or “first” layers refer to those closer to the input, while “deep” or “last” layers are closer to the output. Here, “deep layers” include those around the midpoint, e.g., layers 25–48 in a 48-layer GPT2-XL.