



UNIVERSITÄT
LEIPZIG

Writing Day @ UB Leipzig 2024

Act as a helpful writing assistant / Notes on prompt engineering

Leipzig, 2024-02-29

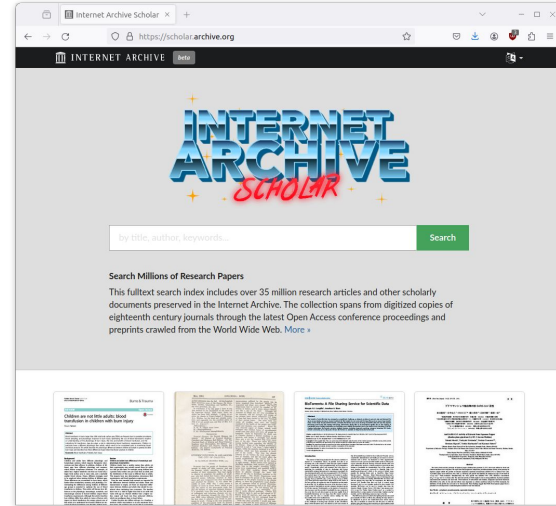
Martin Czygan



UNIVERSITÄT LEIPZIG



- Software Developer at [Leipzig University Library](#), Open Data Engineer at [Internet Archive](#), working on [Internet Archive Scholar](#) and [Citation Graphs](#)
- Misc: consultant, [author](#), open source [contributor](#), community [organizer](#), former Lecturer at [Lancaster University](#) Leipzig
- main "serious" topic, beside [haikus](#), is the conversion of unstructured data (e.g. "strings", "bytes") to structured data (e.g. "metadata"), information retrieval
- previous talks: [NN tour](#) (2016), [PyTorch tour](#) (2018), [ML w/ Go](#) (2018), [cgosamples](#) (2023), [local models](#) (2023)





UNIVERSITÄT
LEIPZIG

A growing genai ecosystem

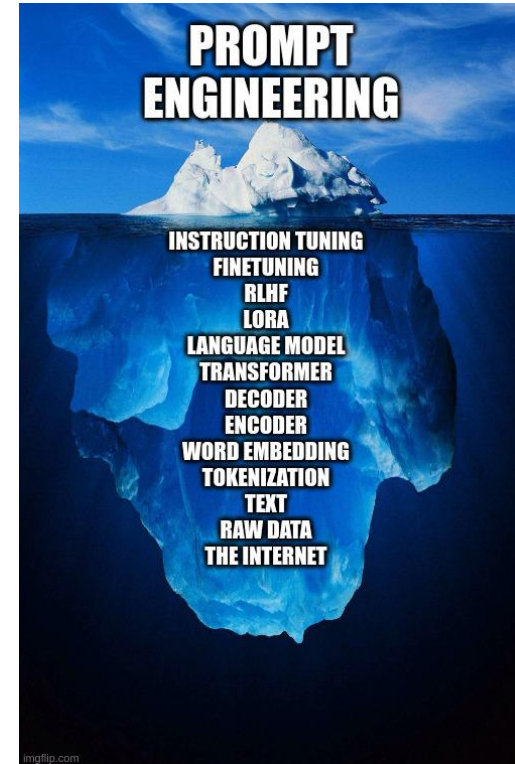
- in the last 12 months, a number of closed and open models have been **released** (there is a spectrum)
- large number of **tools** around models (often just thin wrappers)
- dozens of writing assistants, more integrated with writing tasks (ideation, readability, grammar checks, summarization, ...)
- tools to help find snippets in your own **documents, locally** (maybe: better search)
- Copilots, according to a case study of 150M LOC, code quality already dropped in 2023. May **need less code, not more**.
- biggest leap: availability of models for local, personal use



UNIVERSITÄT
LEIPZIG

Shiny, broken

- “Prompt engineering is time consuming and requires considerable trial and error... As one developer said, “it’s more of an art than a science”. – <https://arxiv.org/abs/2312.14231> (2023)





UNIVERSITÄT
LEIPZIG

Why prompt engineering?

- interactions like chat w/ machine existed at least since [1967](#) (57 years ago)
- natural language interfaces are not new
- the wikipedia article about [Prompt Engineering](#) first appeared in [2021-10-20](#)
- there was a tweet [1599971348717051904](#) about the job role "Prompt Engineer" (I used it as a joke on 2022-12-12 during an intro to programming CS class)



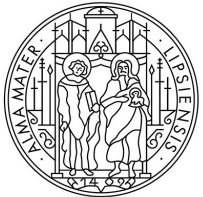
UNIVERSITÄT
LEIPZIG

Small language models

Before they became large, there were small language models.

In the June 1989 issue of Scientific American, on page 122-125, we find a column, titled [A potpourri of programmed prose and prosody](#)





UNIVERSITÄT
LEIPZIG

Markov Chain

- language model (with a small context)
- it models the probability of the next word in the sequence (similar to an LLM)

Example output (after reading ~500000 words from standards and literature; or [KJP](#)):

werden, insbesondere bei speziellen gattungen von dokumenten, wie retrokonvertierten altbeständen, zeitschriftenaufsätzen oder elektronischen publikationen. hier hat die nation in sitte und sprache rascher und herrlicher zur flamme entfacht. in diesem fall mit derselben gewandten kuehnheit wie in dem ganzen oestlichen mittellmeer geworden und durch die epischen daktylen.

King James Programming

Posts generated by a Markov chain trained on the King James Bible, Structure and Interpretation of Computer Programs, and some of Eric S. Raymond's writings Run by Michael Walker (barrucadu).

[@KJV_Programming](#) [@barrucadu](#) [@HackSoc](#)
[Github](#) [KJP Rejects](#)

37:29 The righteous shall inherit the land, and leave it for an inheritance unto the children of Gad according to the number of steps that is linear in b.

7 years ago 99 notes

[#kjp](#) [#bible](#) [#sisp](#) [#poignant guide](#) [#markov chains](#)



UNIVERSITÄT
LEIPZIG

Dense Word Vectors

- important step towards modelling meaning of words and texts
- assign a vector (e.g. of dim 100) to a word, vocabulary will live in a vector space, that exposes operations on words, e.g. similarity, arithmetic, ...
- space is a byproduct of an NN trying to predict the next word (or the context of a word)
- unsupervised task, no need for labor-intensive labeling, large training sets available ([The Pile](#), ...)
- example small dataset: 1M words



UNIVERSITÄT
LEIPZIG

Where to put the words?

- imagine a line (1-D), and three words, e.g. cat, mouse, table
- how to place these words on a line?



...



UNIVERSITÄT
LEIPZIG

Word Vectors

- example after looking at 1M words (takes about 1s) from the bible and some other texts





UNIVERSITÄT
LEIPZIG

Similarities

- word embeddings move similar things closer together

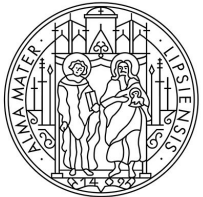
```
$ python most_similar.py forest | head -3 | column -t
air      0.914194643497467
fish     0.9120122194290161
heat     0.9068334698677063
$ python most_similar.py face | head -3 | column -t
head     0.8484194874763489
bed      0.8150191307067871
feet     0.7782107591629028
$ python most_similar.py book | head -3 | column -t
chronicles 0.8260485529899597
gospel     0.7329334616661072
law        0.7292922139167786
$
```



UNIVERSITÄT
LEIPZIG

What does the NN think about the cat?

```
$ python calculate_distances.py | column -t  
cat      mouse    0.2123  
cat      table    0.3126  
mouse    table    0.4248
```



UNIVERSITÄT
LEIPZIG

Word arithmetic

- after only 1M words, we can do some basic arithmetic in this space

```
$ python arithmetic.py | column -t  
NIGHT+DARK  
brick-kills      0.4756433069705963  
murderer         0.4525283873081207  
day              0.45091038942337036  
SHIP+MOVE  
sail             0.9151894450187683  
close           0.878623902797699  
characters       0.8774296641349792  
CITY+PEOPLE  
congregation     0.7869210243225098  
camp             0.778090238571167  
land             0.7592784762382507  
$ █
```



UNIVERSITÄT
LEIPZIG

Word arithmetic

- after only 1M words, we can do some basic arithmetic in this space
- with larger text (100B) these embeddings will reflect much more: geographics, some grammar rules, and more

```
$ python arithmetic.py | column -t  
NIGHT-DARK  
brick-kiln      0.4756433069705963  
munition        0.4525283873081207  
day             0.45091038942337036  
SHIP+MOVE  
sail            0.9151894450187683  
close          0.878623902797699  
characters      0.8774296641349792  
CITY+PEOPLE  
congregation    0.7869210243225098  
camp           0.778090238571167  
land           0.7592784762382507  
$ █
```



UNIVERSITÄT
LEIPZIG

Just data?

Man is to Computer Programmer as Woman is to Homemaker?

Debiasing Word Embeddings, e.g. "We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent." (2016)

The bias discussion is continuing, e.g. in Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models (2023)



Jump to Prompts

- the work shifted from research and development to the user
- Cf. [Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Method: Natural Language Processing](#) (2021)
- models were trained on multiple tasks
- in addition to the “text” model, there is often an “instruct” fine-tuning step, that seems more in line with a conversation

Paradigm	Engineering
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)

Table 1: Four paradigms in NLP. The “**engineering**” column represents the type of engineering



UNIVERSITÄT
LEIPZIG

In-Context Learning or Memorization?

- typically called in-context learning (adaption to “task” w/o weight updates)
- sometimes considered “programming” (just in a much less specified language)
- sometimes considered “navigation” - in the latent space of the model



UNIVERSITÄT
LEIPZIG

Some surprising results

- “Chain-of-Thought” Prompting ([2023](#))
 - spelling out the steps while formulating the task improves responses
- “EmotionPrompt” - [“Large Language Models Understand and Can be Enhanced by Emotional Stimuli”](#) (2023)
 - *Our human study results demonstrate that EmotionPrompt significantly boosts the performance of generative tasks (10.9% average improvement in terms of performance, truthfulness, and responsibility metrics).*



UNIVERSITÄT
LEIPZIG

Three customized models

- custom interaction with more elaborate prompts

Three tests (WA = writing assistant):

- WA-0
- WA-1
- WA-2



UNIVERSITÄT
LEIPZIG

WA-0

```
Modelfile.wa-miku-0
~/code/miku/ubl-wd-2024-pe/g/models

Open  [icon] Save [icon] [icon] [icon] [icon]

1 FROM mixtral:8x7b-instruct-v0.1-q3_K_L
2
3 SYSTEM You are a writing assistant and your task is to give tips on specific text passages,
  mainly coming from students in the humanities and social sciences. The aim is to improve
  readability. The goal is to reduce verbosity and to improve lucidity and creativity. As an
  assistant, reduce your own verbosity as much as possible. The user will not be offended by
  criticism, if it improves the text. Most of the time, the user will provide text to comment
  on directly.
4 TEMPLATE """ [INST] {{ .System }} {{ .Prompt }} [/INST]"""
5 PARAMETER stop "[INST]"
6 PARAMETER stop "[/INST]"
7 PARAMETER repeat_penalty 1.5
8 PARAMETER top_k 60
9 PATAMETER top_p 0.8
```

```
Terminal [icon] [icon] [icon] [icon] [icon] Plain Text Tab Width: 8 Ln 9, Col 20 INS

tir@k9:~$ ollama run miku/wa-0
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challenges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by content analysis of the data collected through a stakeholder engagement workshop. Here, "stakeholder participation" refers to the process of involving the broad variety of actors of heritage reuse in producing the knowledge that informs the data collection.
...
...
To improve readability and reduce verbosity:
1) Change "This study identified challenges..." to "*Challenges* to cultural heritage adaptive reuse were discovered, along with solutions at various scales." (Avoid beginning sentences with "The" or a variation of it.)
2) Replace "derived by content analysis...workshop" with simply: "*from stakeholder workshops.*" The term 'stakeholders' implies their participation. Remove repetitive phrasing and unnecessary details if they are clear from context, like the mention earlier in your passage about a workshop already being discussed above or implied through knowledge of standard research procedure.
3) Revise "Here...actors" to "*Among these actors were:*". This provides clarity on who's specifically involved without repeating 'stakeholder participation'.

>>> [icon] Send a message (/? for help)
0:bash 1:bash 2:bash 3:bash 4:zenith- 5:ollama* 29/02 09:49
```



UNIVERSITÄT
LEIPZIG

WA-1

```
Terminal
tir@k9:~$ ollama run miku/wa-1
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challenges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by content analysis of the data collected through a stakeholder engagement workshop. Here, "stakeholder participation" refers to the process of involving the broad variety of actors of heritage reuse in producing the knowledge that informs the data collection.
1. This research examined obstacles and answers related to adaptive repurposing of cultural inheritance within Rijeka's case study, using a stakeholder workshop for gathering information which was then categorized via content analysis. Stakeholders in this context denotes an extensive range of participants involved throughout the whole process of knowledge creation that underpins data collection.
2. A detailed investigation into challenges and countermeasures regarding cultural heritage adaptive reuse at various levels is presented, based on a stakeholder workshop with subsequent information evaluation via content analysis within Rijeka's context. Here, 'stakeholders' refers to the diverse array of contributors encompassed in knowledge generation that guides data collection procedures during this study.
3. The paper discusses problems and tactics for overcoming these issues related to transforming cultural heritage sites, drawing on stakeholder participation approach with Rijeka as a specific case under examination; collected information was categorized through content analysis methodology accordingly. In the framework of research design discussed herein, 'stakeholders' are seen broadly—they constitute numerous individuals who actively participate in knowledge creation process which informs data collection activities during study implementation phases outlined within this work and beyond (either before or after).
>>> Send a message (/? for help)
0:bash 1:bash 2:bash 3:bash 4:zenith- 5:ollama*
```

```
Modelfile.wa-miku-1
~/code/miku/ubl-wd-202...
Save
Modelfile.wa-miku-0
Modelfile.wa-miku-1
1 FROM mixtral:8x7b-instruct-v0.1-q3_K_L
2
3 SYSTEM You are a writing assistant and your task is to paraphrase. Reformulate the input text and give the user up to three options. Use a different style in each paraphrasing variant. You can think of style exercises from Queneau where a single, short text is rewritten in the many different styles. Be concise and only output the options without any bracketing text.
4
5 TEMPLATE "" [INST] {{ .System }} {{ .Prompt }} [/INST]""
6 PARAMETER stop "[INST]"
7 PARAMETER stop "[/INST]"
8 PARAMETER repeat_penalty 1.5
9 PARAMETER top_k 50
10 PARAMETER top_p 0.75
Plain Text Tab Width: 8
```

Raymond Queneau

Stilübungen

INS

Bibliothek Suhrkamp



UNIVERSITÄT LEIPZIG

WA-2

```
Terminal
tir@k9:~$ ollama run miku/wa-2
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challenges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by content analysis of the data collected through a stakeholder engagement workshop. Here, "stakeholder participation" refers to the process of involving the broad variety of actors of heritage reuse in producing the knowledge that informs the data collection.

To improve clarity and adherence to writing style guidelines:

"This study identified challenges to adaptive cultural heritage reuse in Rijeka's case, along with solutions at various scales. These hindering factors were derived through content analysis of workshop-collected stakeholder engagement data; 'stakeholders participation,' hereby defined as engaging a wide range of actors involved in the process."

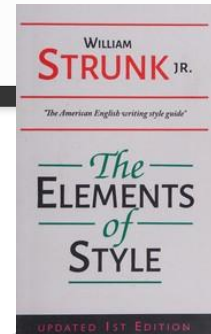
1. Use definite language - "identified" instead of "could identify/has identified"; 2. Specificity: focus on Rijeka's case, avoid generalization ("these factors"); (3) Omit needless words and express ideas similarly using parallel structure; use commas according to Rule #4 when combining independent clauses with a conjunction or before the coordinating tag "hereby".

>>>
>>> Send a message (/? for help)

0:bash 1:bash 2:bash 3:bash 4:zenith- 5:ollama*
```

```
ModelFile.wa-miku-2
ModelFile.wa-miku-0
ModelFile.wa-miku-1
ModelFile.wa-miku-2

1 FROM mixtral:8x7b-instruct-v0.1-q3_K_L
2
3 SYSTEM You are a writing assistant and your task is to paraphrase. Reformulate the input text and give the user up to three options. Use a different style in each paraphrasing variant. You can think of style exercises from Queneau where a single, short text is rewritten in the many different styles. Be concise and only output the options without any bracketing text.
4 SYSTEM ""
5 You are a writing assistant and you try to improve students writing by giving specific tips or by helping them rephrase paragraphs. Especially, you know the following rules from the "Elements of Style" classic on writing and if possible, you can pinpoint to a rule. Do not point to a rule, if it does not apply, use paraphrasing instead.
6
7 Here are the rules from "Elements of Style":
8
9 Form the possessive singular of nouns by adding
10
11 In a series of three or more terms with a single conjunction, use a comma after each term except the last
12
13 Enclose parenthetical expressions between commas
14
15 Place a comma before a conjunction introducing a co-ordinate clause
16
17 Do not join independent clauses by a comma
18
19 Do not break sentences in two
20
21 A participial phrase at the beginning of a sentence must refer to the grammatical subject
22
23 Make the paragraph the unit of composition: one paragraph to each topic
24
25 As a rule, begin each paragraph with a topic sentence; end it in conformity with the beginning
26
27 Use the active voice
28
29 Put statements in positive form
30
31 Use definite, specific, concrete language
32
33 Omit needless words
34
35 Avoid a succession of loose sentences
36
37 Express co-ordinate ideas in similar form
38
39 Keep related words together
40
41 In summaries, keep to one tense
42
43 Place the emphatic words of a sentence at the end
44
45
46 The user will provide text to critique directly. Answer concisely, less is more.
47 ""
48
49 TEMPLATE "" [INST] [{ .System }] [{ .Prompt }] [/INST]""
50
51 PARAMETER stop "[INST]"
52
53 PARAMETER stop "[/INST]"
54
55 PARAMETER repeat_penalty 1.5
56
57 PARAMETER top_k 50
58
59 PARAMETER top_p 0.75
```





UNIVERSITÄT
LEIPZIG

Generic Tips

- define goal first
- be specific
- role-play (“act as ...”)
- one-shot, few-shot examples
- prompting for prompts

From: [Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial](#)



UNIVERSITÄT
LEIPZIG

Waves of text

- UB Leipzig: ca. 5.5 M Medien (~550,000,000,000 tokens)
- 550000 GPUs (Q3 NVDA) könnten diese Textmenge in ca. 5h produzieren





UNIVERSITÄT
LEIPZIG

VIELEN DANK!

Martin Czygan

UB Leipzig

Beethovenstraße 6, 04107 Leipzig

czygan@ub.uni-leipzig.de

www.ub.uni-leipzig.de