UNIVERSITÄT
LEIPZIG

Writing Day @ UB Leipzig 2024

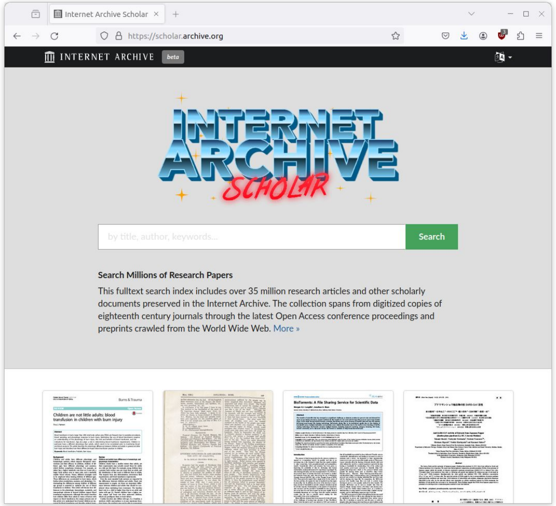
Act as a helpful writing assistant / Notes on prompt engineering

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)



Search Millions of Research Papers

This fulltext search index includes over 35 million research articles and other scholarly documents preserved in the Internet Archive. The collection spans from digitized copies of eighteenth century journals through the latest Open Access conference proceedings and preprints crawled from the World Wide Web. More »

UNIVERSITÄT
LEIPZIG

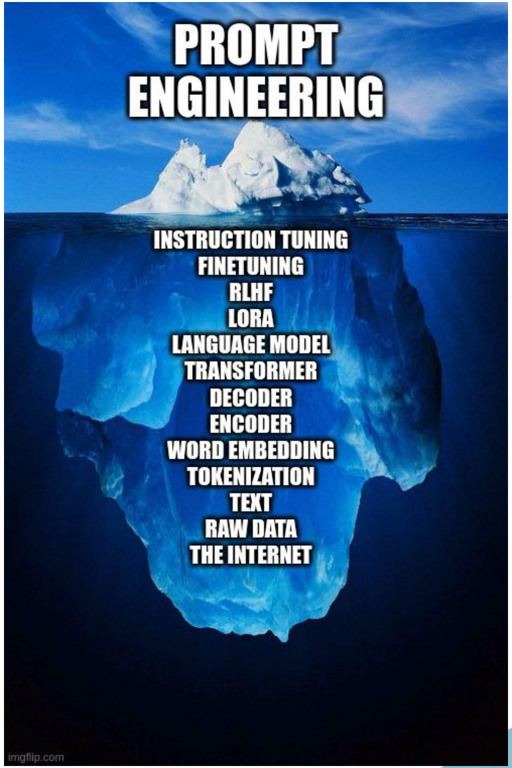
A growing genai ecosystem

- 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)



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)

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]

COMPUTER RECREATIONS

A potpourri of programmed prose and prosody

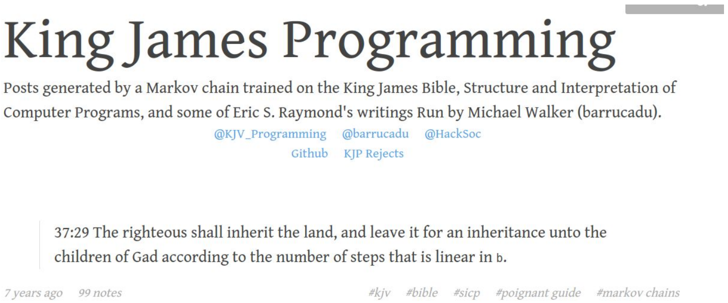
by A. K. Dewdney

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 mittelmeer 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.

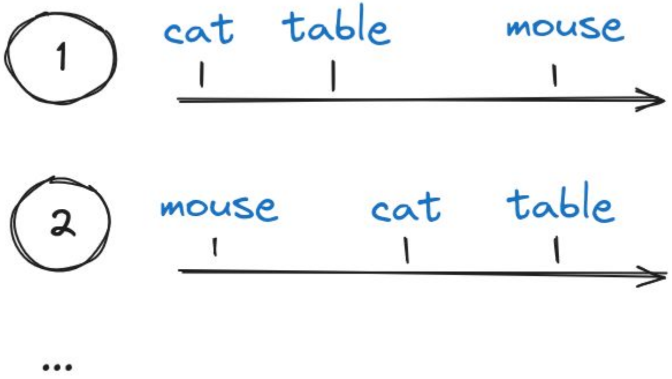
7 years ago 99 notes #kjv #bible #sicp #poignant guide #markov chains

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

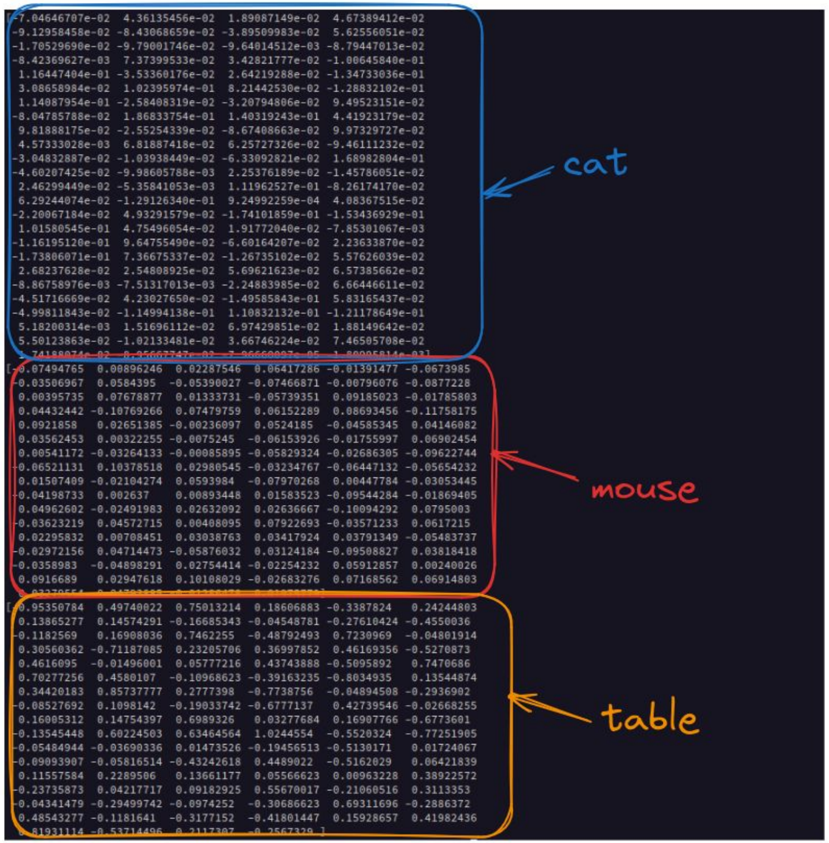
Where to put the words?

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



Word Vectors

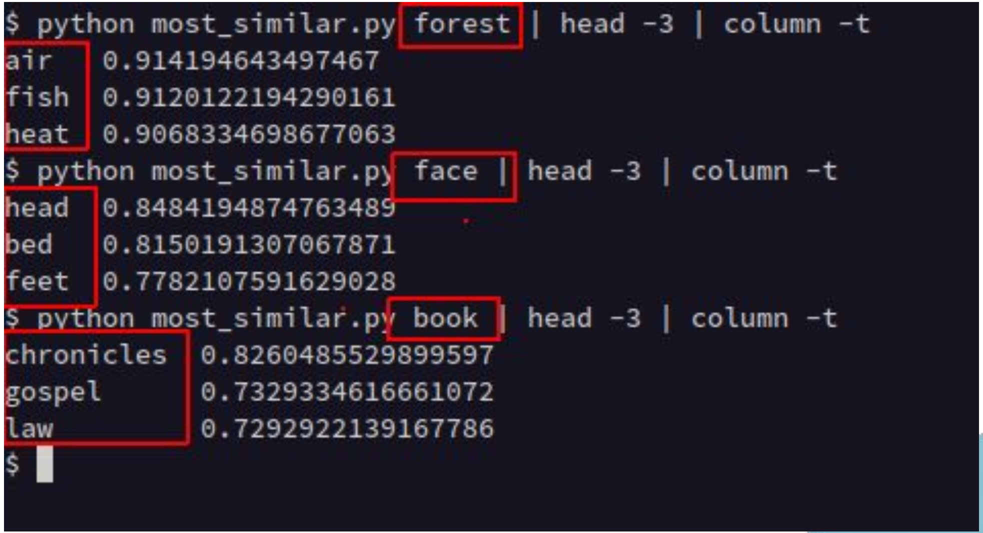
- example after looking at 1M words (takes about 1s) from the bible and some other texts



```
[-7.04646707e-02  4.36135456e-02  1.89087149e-02  4.67389412e-02
 -9.12958458e-02 -8.43068659e-02 -3.89509983e-02  5.62556051e-02
 -1.70529690e-02 -9.79001746e-02 -9.64014512e-03 -8.79447013e-02
 -8.42369627e-03  7.37399533e-02  3.42821777e-02 -1.00645840e-01
  1.16447404e-01 -3.53360176e-02  2.64219288e-02 -1.34733036e-01
  3.08658984e-02  1.02395974e-01  8.21442530e-02 -1.28832102e-01
  1.14087954e-01 -2.58408319e-02 -3.20794806e-02  9.49523151e-02
 -8.04785788e-02  1.86833754e-01  1.40319243e-01  4.41923179e-02
  9.81888175e-02 -2.55254339e-02 -8.67408663e-02  9.97329727e-02
  4.57333028e-03  6.81887418e-02  6.25727326e-02 -9.46111232e-02
 -3.04832887e-02 -1.03938449e-02 -6.33092821e-02  1.68982804e-01
 -4.60207425e-02 -9.98605788e-03  2.25376189e-02 -1.45786051e-02
  2.46299449e-02 -5.35841053e-03  1.11962527e-01 -8.26174170e-02
  6.29244074e-02 -1.29126340e-01  9.24992259e-04  4.08367515e-02
 -2.20067184e-02  4.93291579e-02 -1.74101859e-01 -1.53436929e-01
  1.01580545e-01  4.75496054e-02  1.91772040e-02 -7.85301067e-03
 -1.16195120e-01  9.64755490e-02 -6.60164207e-02  2.23633870e-02
 -1.73806071e-01  7.36675337e-02 -1.26735102e-02  5.57626039e-02
  2.68237628e-02  2.54808925e-02  5.69621623e-02  6.57385662e-02
 -8.86758976e-03 -7.51317013e-03 -2.24883985e-02  6.66446611e-02
 -4.51716669e-02  4.23027650e-02 -1.49585843e-01  5.83165437e-02
 -4.99811843e-02 -1.14994138e-01  1.10832132e-01 -1.21178649e-01
  5.18200314e-03  1.51696112e-02  6.97429851e-02  1.88149642e-02
  5.50123863e-02 -1.02133481e-02  3.66746224e-02  7.46505708e-02
  1.74188074e-02  8.95667747e-02 -7.96660087e-05  1.80095814e-03]
[-0.07494765  0.00896246  0.02287546  0.06417286 -0.01391477 -0.0673985
 -0.03506967  0.0584395  -0.05390027 -0.07466871 -0.00796076 -0.0877228
  0.00395735  0.07678877  0.01333731 -0.05739351  0.09185023 -0.01785803
  0.04432442 -0.10769266  0.07479759  0.06152289  0.08693456 -0.11758175
  0.0921858   0.02651385 -0.00236097  0.0524185  -0.04585345  0.04146082
  0.03562453  0.00322255 -0.0075245  -0.06153926 -0.01755997  0.06902454
  0.00541172 -0.03264133 -0.00085895 -0.05829324 -0.02686305 -0.09622744
 -0.06521131  0.10378518  0.02980545 -0.03234767 -0.06447132 -0.05654232
  0.01507409 -0.02104274  0.0593984  -0.07970268  0.00447784 -0.03053445
 -0.04198733  0.002637    0.00893448  0.01583523 -0.09544284 -0.01869405
  0.04962602 -0.02491983  0.02632092  0.02636667 -0.10094292  0.0795003
 -0.03623219  0.04572715  0.00408095  0.07922693 -0.03571233  0.0617215
  0.02295832  0.00708451  0.03038763  0.03417924  0.03791349 -0.05483737
 -0.02972156  0.04714473 -0.05876032  0.03124184 -0.09508827  0.03818418
 -0.0358983  -0.04898291  0.02754414 -0.02254232  0.05912857  0.00240026
  0.0916689   0.02947618  0.10108029 -0.02683276  0.07168562  0.06914803
  0.03270554 [illegible] ]
[ 0.95350784  0.49740022  0.75013214  0.18606883 -0.3387824   0.24244803
  0.13865277  0.14574291 -0.16685343 -0.04548781 -0.27610424 -0.4550036
 -0.1182569   0.16908036  0.7462255  -0.48792493  0.7230969  -0.04801914
  0.30560362 -0.71187085  0.23205706  0.36997852  0.46169356 -0.5270873
  0.4616095  -0.01496001  0.05777216  0.43743888 -0.5095892   0.7470686
  0.70277256  0.4580107  -0.10968623 -0.39163235 -0.8034935   0.13544874
  0.34420183  0.85737777  0.2777398  -0.7738756  -0.04894508 -0.2936902
 -0.08527692  0.1098142  -0.19033742 -0.6777137   0.42739546 -0.02668255
  0.16005312  0.14754397  0.6989326   0.03277684  0.16907766 -0.6773601
 -0.13545448  0.60224503  0.63464564  1.0244554  -0.5520324  -0.77251905
 -0.05484944 -0.03690336  0.01473526 -0.19456513 -0.5130171   0.01724067
 -0.09093907 -0.05816514 -0.43242618  0.4489022  -0.5162029   0.06421839
  0.11557584  0.2289506   0.13661177  0.05566623  0.00963228  0.38922572
 -0.23735873  0.04217717  0.09182925  0.55670017 -0.21060516  0.3113353
 -0.04341479 -0.29499742 -0.0974252  -0.30686623  0.69311696 -0.2886372
  0.48543277 -0.1181641  -0.3177152  -0.41801447  0.15928657  0.41982436
  0.81931114 -0.53714496  0.2117307  -0.2567329 ]
```

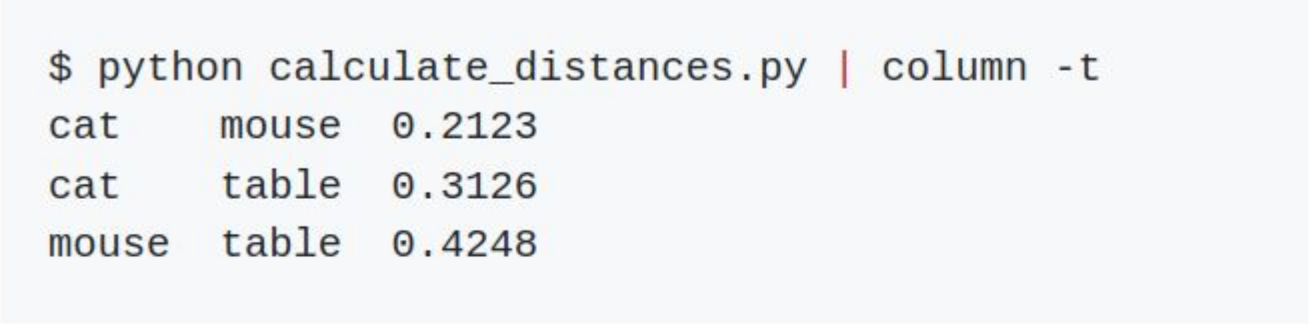
Similarities

- 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
$
```

What does the NN think about the cat?

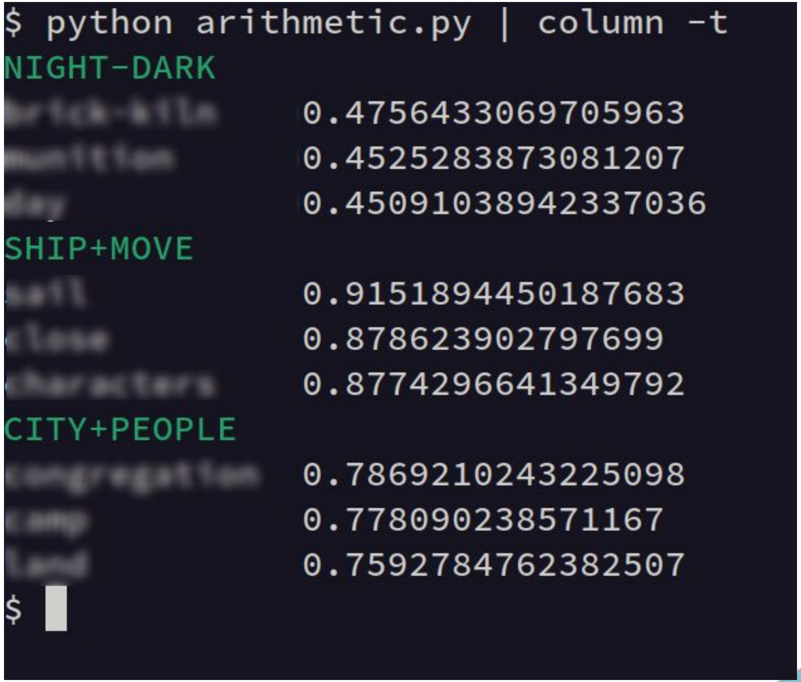


```
$ python calculate_distances.py | column -t
cat    mouse  0.2123
cat    table  0.3126
mouse  table  0.4248
```

UNIVERSITÄT
LEIPZIG

Word arithmetic

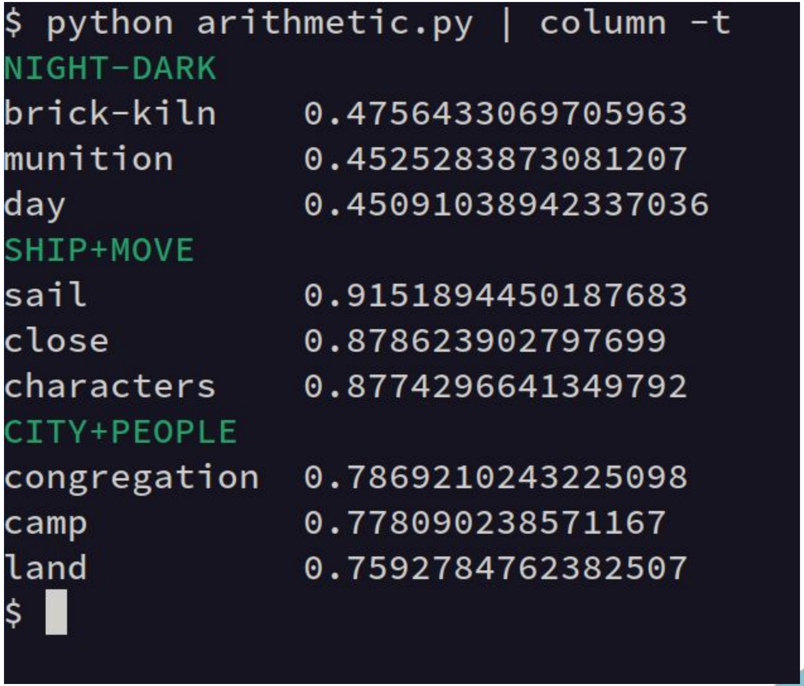
- after only 1M words, we can
  do some basic arithmetic in this space



```
$ python arithmetic.py | column -t
NIGHT-DARK
[redacted]     0.4756433069705963
[redacted]     0.4525283873081207
[redacted]     0.45091038942337036
SHIP+MOVE
[redacted]     0.9151894450187683
[redacted]     0.878623902797699
[redacted]     0.8774296641349792
CITY+PEOPLE
[redacted]     0.7869210243225098
[redacted]     0.778090238571167
[redacted]     0.7592784762382507
$ █
```

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

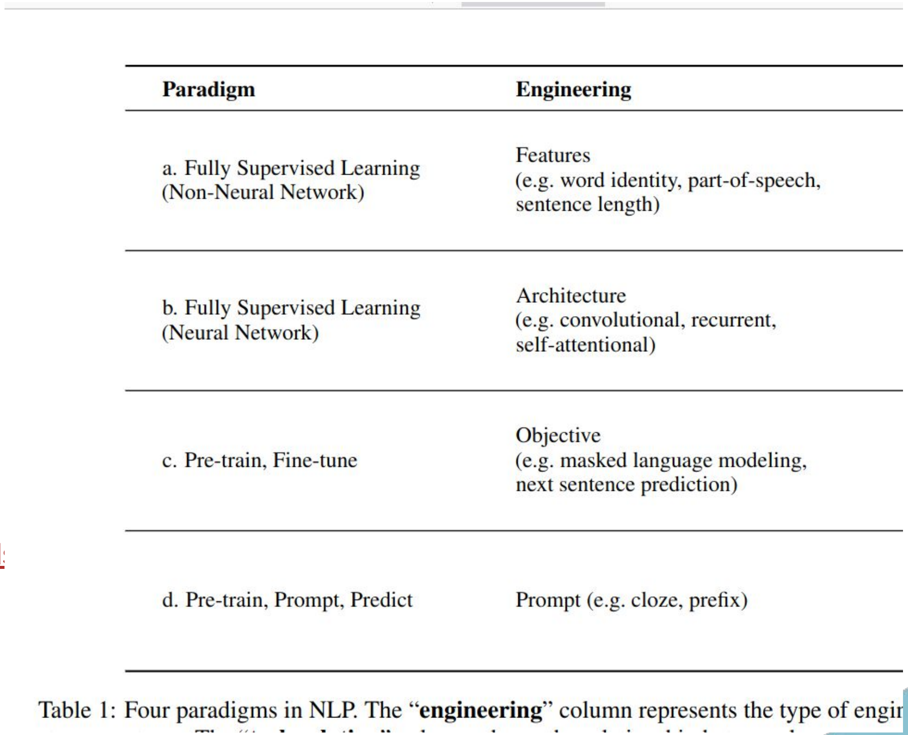
Just data?

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 engir

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

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

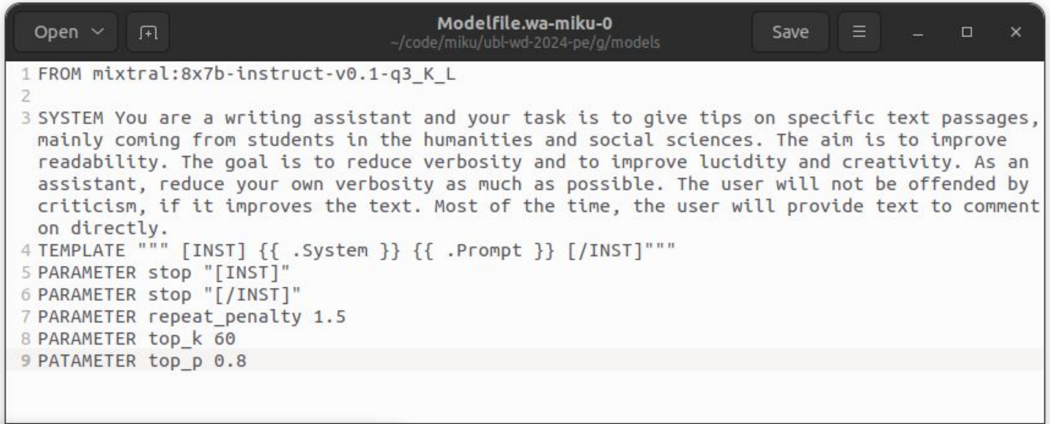
- custom interaction with more elaborate prompts

Three tests (WA = writing assistant):

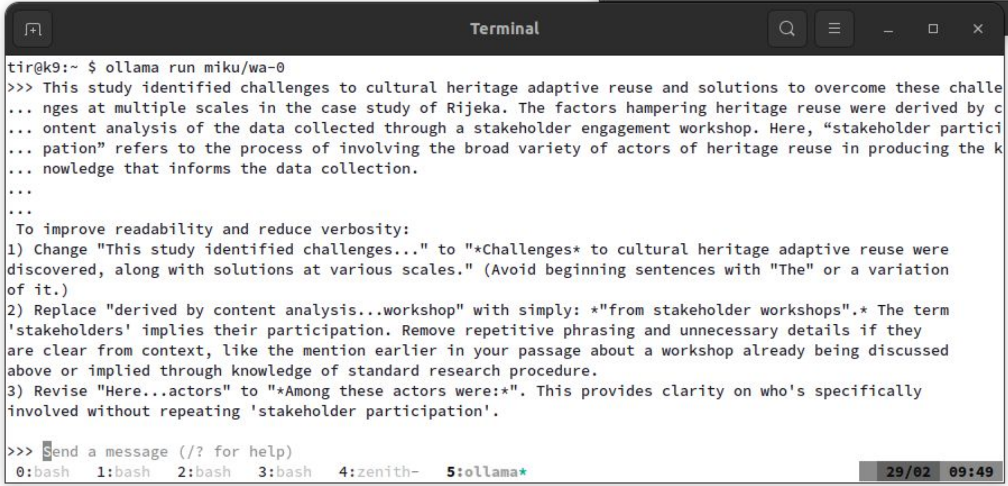
- WA-0
- WA-1
- WA-2

UNIVERSITÄT
LEIPZIG

WA-0



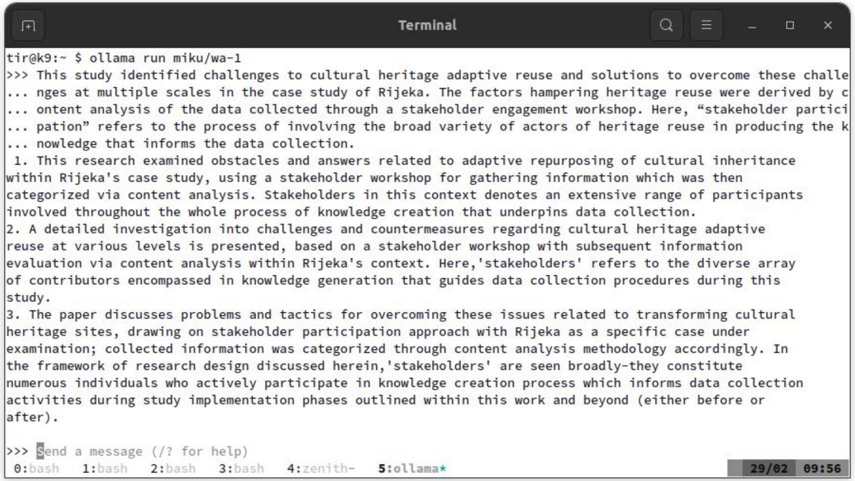
```
FROM mixtral:8x7b-instruct-v0.1-q3_K_L

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.
TEMPLATE """ [INST] {{ .System }} {{ .Prompt }} [/INST]"""
PARAMETER stop "[INST]"
PARAMETER stop "[/INST]"
PARAMETER repeat_penalty 1.5
PARAMETER top_k 60
PATAMETER top_p 0.8
```



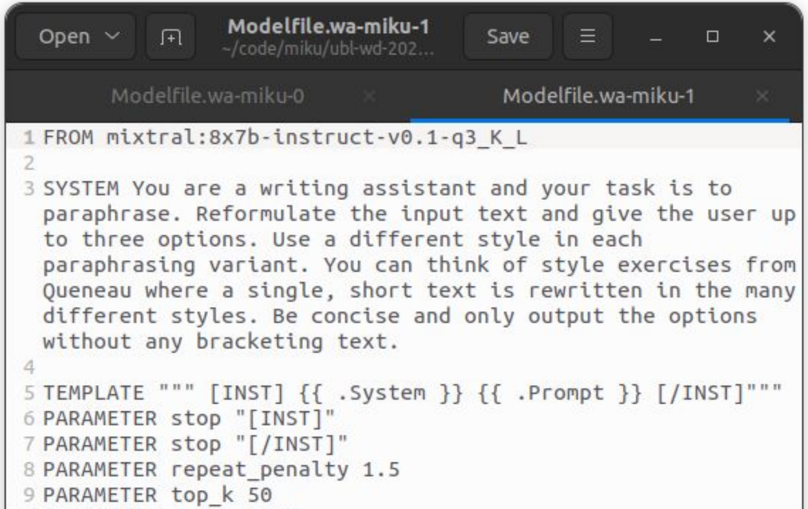
```
tir@k9:~ $ ollama run miku/wa-0
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challe
... nges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by c
... ontent analysis of the data collected through a stakeholder engagement workshop. Here, “stakeholder partici
... pation” refers to the process of involving the broad variety of actors of heritage reuse in producing the k
... nowledge 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'.

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

UNIVERSITÄT
LEIPZIG

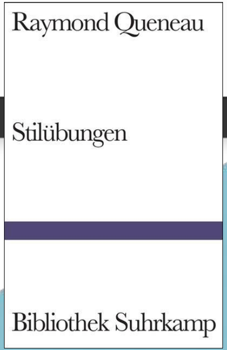
WA-1



```
tir@k9:~ $ ollama run miku/wa-1
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challe
... nges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by c
... ontent analysis of the data collected through a stakeholder engagement workshop. Here, “stakeholder partici
... pation” refers to the process of involving the broad variety of actors of heritage reuse in producing the k
... nowledge 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*                                    29/02  09:56
```



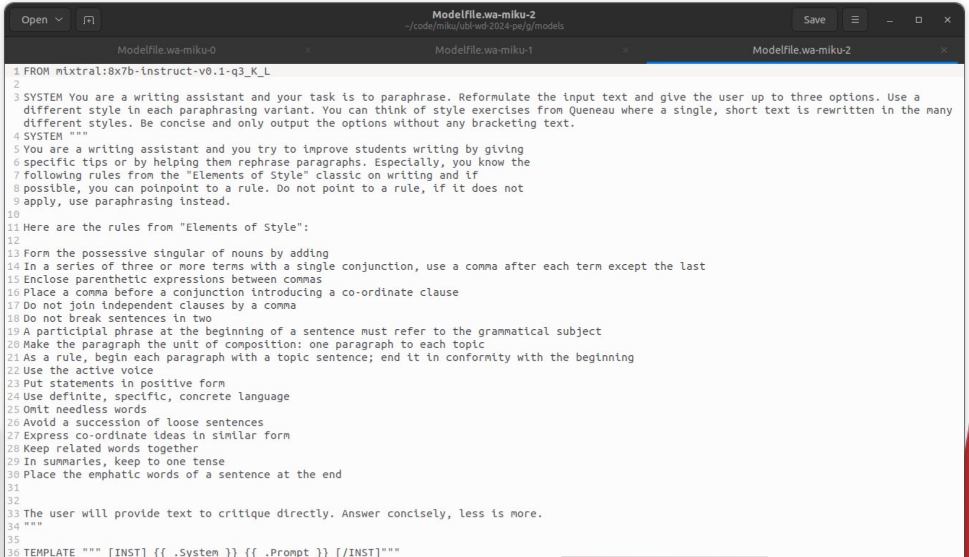
```
FROM mixtral:8x7b-instruct-v0.1-q3_K_L

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.

TEMPLATE """ [INST] {{ .System }} {{ .Prompt }} [/INST]"""
PARAMETER stop "[INST]"
PARAMETER stop "[/INST]"
PARAMETER repeat_penalty 1.5
PARAMETER top_k 50
PATAMETER top_p 0.75
```


Raymond Queneau

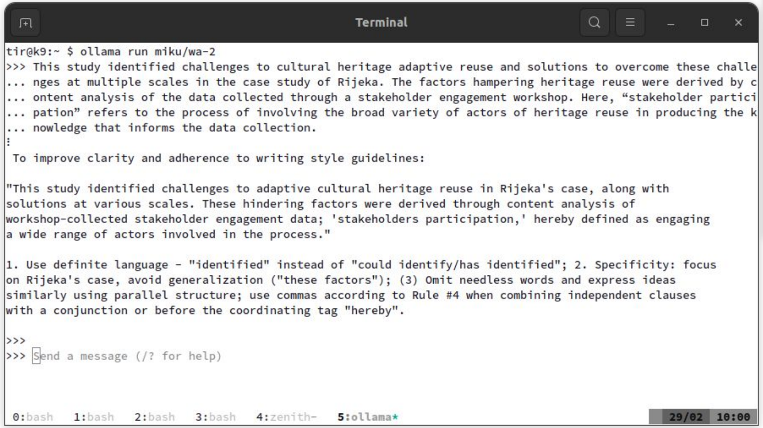
Stilübungen

Bibliothek Suhrkamp

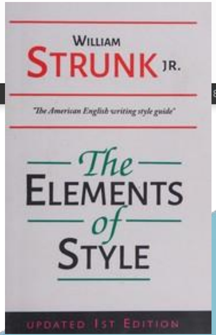
WA-2



```
FROM mixtral:8x7b-instruct-v0.1-q3_K_L

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.
SYSTEM """
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 poinpoint to a rule. Do not point to a rule, if it does not
apply, use paraphrasing instead.

Here are the rules from "Elements of Style":

Form the possessive singular of nouns by adding
In a series of three or more terms with a single conjunction, use a comma after each term except the last
Enclose parenthetic expressions between commas
Place a comma before a conjunction introducing a co-ordinate clause
Do not join independent clauses by a comma
Do not break sentences in two
A participial phrase at the beginning of a sentence must refer to the grammatical subject
Make the paragraph the unit of composition: one paragraph to each topic
As a rule, begin each paragraph with a topic sentence; end it in conformity with the beginning
Use the active voice
Put statements in positive form
Use definite, specific, concrete language
Omit needless words
Avoid a succession of loose sentences
Express co-ordinate ideas in similar form
Keep related words together
In summaries, keep to one tense
Place the emphatic words of a sentence at the end


The user will provide text to critique directly. Answer concisely, less is more.
"""

TEMPLATE """ [INST] {{ .System }} {{ .Prompt }} [/INST]"""
PARAMETER stop "[INST]"
PARAMETER stop "[/INST]"
PARAMETER repeat_penalty 1.5
PARAMETER top_k 50
PATAMETER top_p 0.75
```



```
tir@k9:~ $ ollama run miku/wa-2
>>> This study identified challenges to cultural heritage adaptive reuse and solutions to overcome these challe
... nges at multiple scales in the case study of Rijeka. The factors hampering heritage reuse were derived by c
... ontent analysis of the data collected through a stakeholder engagement workshop. Here, “stakeholder partici
... pation” refers to the process of involving the broad variety of actors of heritage reuse in producing the k
... nowledge 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)
```



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 media (or ~550,000,000,000 tokens)
- 550000 GPUs (sold by Q3 NVDA) could produce the amount of text in this library in about five hours

UNIVERSITÄT
LEIPZIG

VIELEN DANK!

Martin Czygan
UB Leipzig

Beethovenstraße 6, 04107 Leipzig

czygan@ub.uni-leipzig.de
www.ub.uni-leipzig.de