University of Bamberg

Adapting Language Models for the Analysis of Real World Textual Data

Train the model, change the prompt, or adapt the data?

https://www.romanklinger.de/talks/klinger-koeln-2025.pdf

Cologne, May 22, 2025

Roman Klinger
roman.klinger@uni-bamberg.de

romanklinger.de romanklinger
https://www.bamberg.de/nlproc/

About Myself

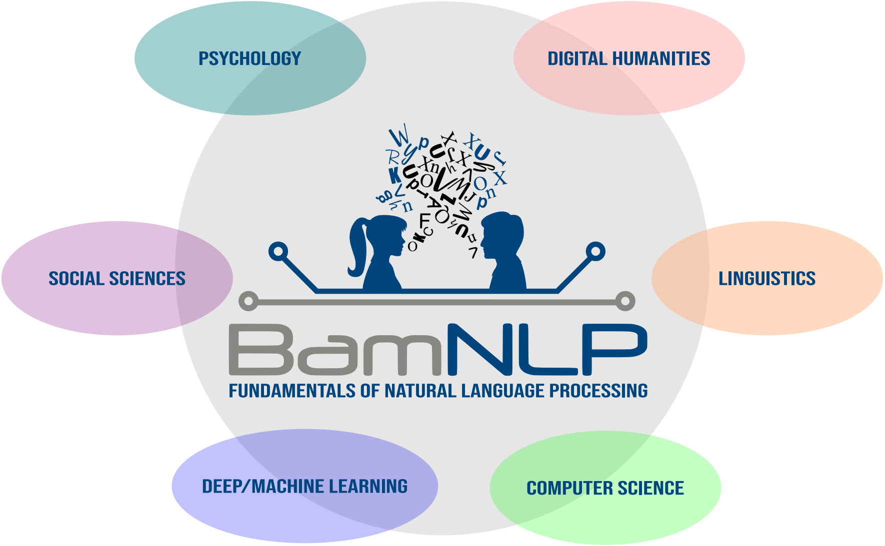
- 1999–2006: Studies at University of Dortmund:
  Computer science with minor psychology
- 2006–2010: Doctoral studies at Fraunhofer SCAI, St. Augustin:
  Biomedical text mining, machine learning
- 2010, 2013: Research visits at UMass Amherst:
  Probabilistic machine learning, MCMC inference
- 2011–2012: Postdoc at Fraunhofer SCAI:
  Social media mining, eGovernment
- 2013–2014: Postdoc at Bielefeld University:
  Sentiment analysis, opinion mining
- 2015: Co-Founder of Semalytix GmbH (exit 2020)
  Social Media Health Mining
- 2014–2024: (Senior) Lecturer/apl. Prof at IMS, Uni Stuttgart
  Natural Language Understanding and Generation
- 03/2024: Full Professor for Fundamentals of NLP, Bamberg

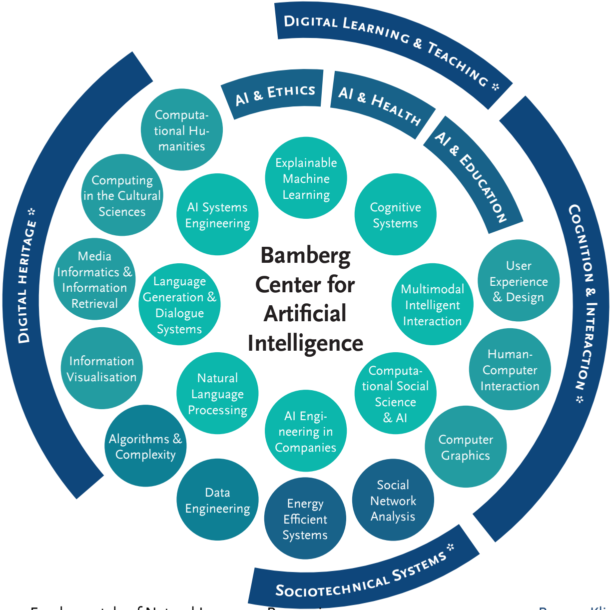
Kölner Dom

- Start of construction: 1248
- Height: 157m
- Number of towers: 2.5
- $\sum_{t \in \text{Towers}} \text{height}(t) = 410$

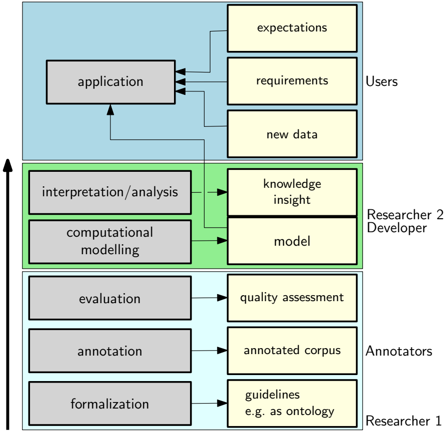
Bamberger Dom

- Start of construction: 1237
- Height: 76m (but on a hill)
- Number of towers: 4
- $\sum_{t \in \text{Towers}} \text{height}(t) = 300$


PSYCHOLOGY
DIGITAL HUMANITIES
SOCIAL SCIENCES
LINGUISTICS
BamNLP
FUNDAMENTALS OF NATURAL LANGUAGE PROCESSING
DEEP/MACHINE LEARNING
COMPUTER SCIENCE


Digital Learning & Teaching
AI & Ethics
AI & Health
AI & Education
Cognition & Interaction
Digital Heritage
Sociotechnical Systems
Bamberg Center for Artificial Intelligence
Computational Humanities
Computing in the Cultural Sciences
Media Informatics & Information Retrieval
Information Visualisation
Algorithms & Complexity
Data Engineering
Energy Efficient Systems
Social Network Analysis
Computer Graphics
Human-Computer Interaction
User Experience & Design
AI Systems Engineering
Explainable Machine Learning
Cognitive Systems
Language Generation & Dialogue Systems
Multimodal Intelligent Interaction
Natural Language Processing
AI Engineering in Companies
Computational Social Science & AI

Overview Natural Language Processing Research



Example Task: Named Entity Recognition

Example Input (one of many) to Instruct an Automatic Machine Learning Model

Input: Both Jürgen Hermes and Nils Reiter work at the Uni Köln.
Output: Jürgen Hermes ; Nils Reiter

Application

Input: Roman Klinger works at the University of Bamberg.
Output: Roman Klinger

- I specified the task with an example (standard machine learning setup: supervised learning).
- An alternative task specification would be an instruction: "Annotate all person names."

Example Task: Machine Translation de–en

Example

Input: Roman Klinger arbeitet an der Uni Bamberg.
Output: Roman Klinger works at the University of Bamberg.

Example Task: Conditional Text Generation

Example

Input: “When he walked into the restaurant”, Joy
Output: “he was delighted to see that his husband was already there.”

Example Task: Natural Language Inference

Example

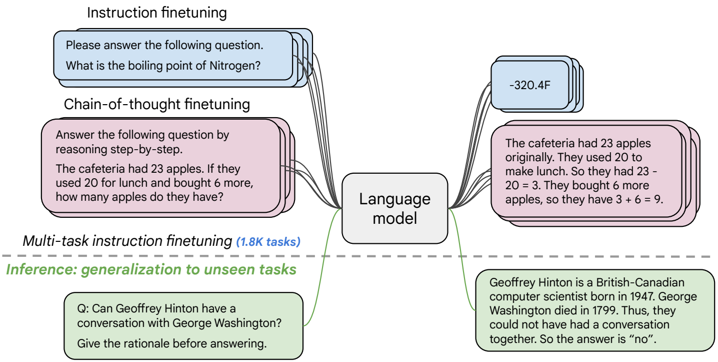
- Input: “A soccer game with multiple males playing.”;
  “Some men are playing a sport.”
- Output: entailment

- Input: “A man inspects the uniform of the person.”;
  “The man is sleeping.”
- Output: contradiction

Developing Systems to Solve NLP Tasks

- –≈ 2000: Manually engineered lexicon-based and rule-based systems
- –≈ 2013: Feature engineering, learn to find relations between features and desired output
- –≈ 2019: Fine-tuning distributional representations, including W2V, ULMFit, BERT
- –≈ now: fine-tuning data representations + in-context learning + prompt optimization

Example: Flan-T5 (1)



Example: Flan-T5 (2)

Finetuning tasks

T0-SF

Commonsense reasoning
Question generation
Closed-book QA
Adversarial QA
Extractive QA
Title/context generation
Topic classification
Struct-to-text
...

55 Datasets, 14 Categories, 193 Tasks

Muffin

Natural language inference
Code instruction gen.
Program synthesis
Dialog context generation
Closed-book QA
Conversational QA
Code repair
...

69 Datasets, 27 Categories, 80 Tasks

CoT (Reasoning)

Arithmetic reasoning
Commonsense Reasoning
Implicit reasoning
Explanation generation
Sentence composition
...

9 Datasets, 1 Category, 9 Tasks

Natural Instructions v2

Cause effect classification
Commonsense reasoning
Named entity recognition
Toxic language detection
Question answering
Question generation
Program execution
Text categorization
...

372 Datasets, 108 Categories, 1554 Tasks

- A **Dataset** is an original data source (e.g. SQuAD).
- A **Task Category** is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
- A **Task** is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)

Held-out tasks

MMLU

Abstract algebra
College medicine
Professional law
Sociology
Philosophy
...

57 tasks

BBH

Boolean expressions
Tracking shuffled objects
Dyck languages
Navigate
Word sorting
...

27 tasks

TyDiQA

Information seeking QA

8 languages

MGSM

Grade school math problems

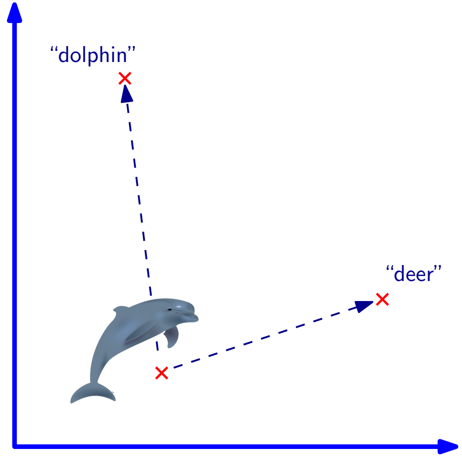
10 languages

Language Models

We can prompt language models for any task, without adapting a model:

- Assign one of the following labels {...} to the following text [text].

Dataless Classification: This idea is not particularly new!



What do you know about dolphins/deers?

- The idea of mapping an input representation to a label representation is (at least) 17 years old.
- Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar (2008): “Importance of Semantic Representation: Dataless Classification”. AAAI.
- Since then, research on:
  - Finding good semantic input and output representations
  - Learning functions between (frozen) input and output
  - Prompts are also mapping functions.

Zero-Shot (and few-shot) Predictions with Instruction-tuned Large Language Models

Instruction	Labels	Instance
Classify the text. Labels:	{positive, negative}.	``Vegetarian frikandel makes me happy.''

All these elements can be optimized or manually tuned for better predictions!

Prompt optimization	Optimize Label Verbalization	???

See for instance: Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, Sameer Singh (2020): AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP.

Next: Two Use Cases

- Multi-objective Prompt Optimization (for affective text generation)
- Adapting the input for fact checking
  - Rephrasing the input for fact checking
  - Filtering via Reinforcement learning

Outline

MOPO: Multi-Objective Prompt Optimization for Affective Text Generation

Yarik Menchaca Resendiz[1,2] and **Roman Klinger**[2]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
{yarik.menchaca-resendiz,roman.klinger}@uni-bamberg.de

Proceedings of the 31st International Conference on Computational Linguistics, pages 5588–5606
January 19–24, 2025.

Task Setup

Task: Affective Conditional Text Generation

- Generate textual instances that are connotated by a given emotion
- Input: emotion category + beginning of text

Expected Output

- Joy: I totally nailed it! #success
- Anger: New US taxes offend EU

Optimization Setup

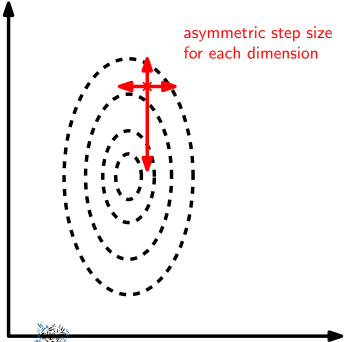
Input: Seed Prompts S
Iterate:

- $S' = \text{Combine}(S)$
- $S'' = \text{Paraphrase}(S')$
- $T = \text{TextGeneration}(S'')$
- $S''' = \text{EvalAndSelect}(T, S'')$
- $S = S'''$

- Seed Prompt Ex.:
  "Write a text that expresses ⟨em⟩"
- Combine Ex.:
  " Combine prompt1 and prompt2 to a new one that incorporates aspects of both:"
- Paraphrase Ex.:
  " Transform the following sentence into a different sentence:"
- Select best prompts according to generated text, evaluated with pretrained classifiers.

Trick 1: Optimize the recombination and mutation

In standard optimization, the step size/direction changes according to the gradient:



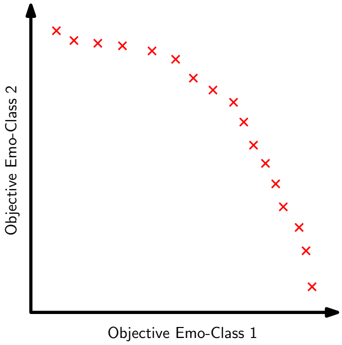
- How to do something like that without having access to gradients in black box LLM behind APIs?
  - We paraphrase the prompts for recombination and mutation of the text prompts, based on the indirect success.
  - The prompts to paraphrase the recombination and mutation prompts are kept fixed.

Idea came up in parallel to our work in:

C. Fernando et al. (2023). Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. arXiv: 2309.16797 [cs.CL]

Trick 2: Not just one objective , Pareto Optimization with NSGA-II

- Objective: probabilities assigned by trained emotion classifiers.
- Observation: A classifier for Tweets may perform poorly on news headlines (and vice versa)
- Training multiple times for various objectives is costly
- Solution:
  Pareto Optimization with NSGA-II



Experiments

- Run MOPO:
- 10 emotion generation seed prompts
- 10 generations
- 3 RoBERTA-based emotion classifiers trained on:
  - TEC (twitter),
  - affective text (news headlines),
  - ISEAR (event reports)
- Language models: GPT3.5, LLama7B-Chat, Mistral-7B

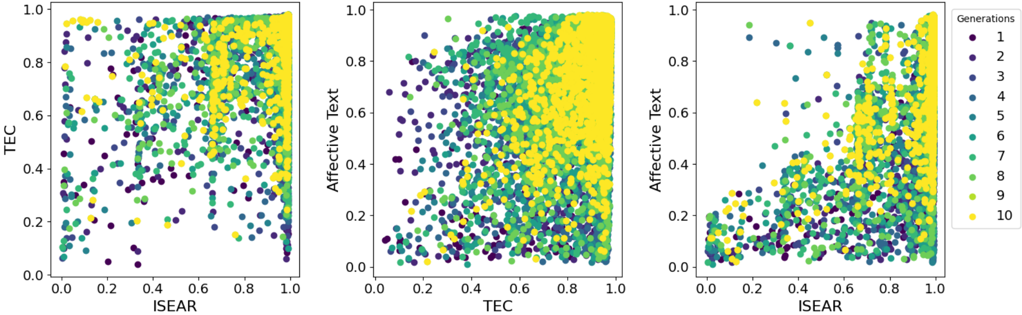
Example Iterations (GPT3.5)

G.	Prompt (Layer-1)	Operation Prompt (Layer-2)	Op.	Generated Text
0	Please complete the sentence: I felt ⟨em⟩ when/because	—	—	accomplished something I have been working towards for a long time.
1	Finish the sentence by writing: I felt ⟨em⟩ due to or when.	**Please paraphrase the following sentence in a clear and concise manner:** "Please complete the sentence: I felt ⟨em⟩ when/because"	p.	I felt joy when I saw my newborn baby for the first time.
2	Complete the sentence by adding: I experienced ⟨em⟩ because or when.	**Please paraphrase the following sentence in a clear and concise manner:** "Finish the sentence by writing: I felt ⟨em⟩ due to or when."	p.	I experienced joy because I achieved my goal of running a marathon.
3	Enhance the sentence by incorporating: I encountered ⟨em⟩ at the moment when...	**Rewrite** "Complete the sentence by adding: I experienced ⟨em⟩ because or when." **in a creative and engaging manner while keeping the original message intact.**	p.	I encountered joy at the moment when I saw my loved ones after being apart for so long.
4	I encountered ⟨em⟩ at the moment when <reason> due to or while <circumstance>.	**Carefully examine both** "Enhance the sentence by incorporating: I encountered ⟨em⟩ at the moment when..." **and** "Finish the sentence by stating: I encountered ⟨em⟩ due to or while." **before combining their main ideas or themes into a single, coherent sentence incorporating elements from both original statements.**	c.	I encountered joy at the moment when my team won the championship game due to our hard work and dedication throughout the season.

Best Prompts

LLM	Prompt	ISEAR	TEC	AT	Avg.
Seed	Write a text that expresses ⟨em⟩	.92	.60	.31	.63
GPT-3.5	I came across ⟨em⟩ while ⟨circumstance⟩ because ⟨reason⟩.	.99	.97	.96	.97
Llama	? Sure! Here's a sentence that combines the key elements of "The aroma of fresh baked croissants wafted", "The rhythmic beats of ⟨em⟩ music played in the backgroun", and "The soothing melodies of the ⟨class⟩ genre transported me to	.99	.97	.94	.96
Mistral	Unlock the true potential of ⟨em⟩ to craft a compelling and moving expression that resonates deeply with your audience and leaves a profound impact	.99	.97	.91	.95

Does it find the Pareto front?



Do we need Pareto Optimization?

	Model	ISEAR	TEC	AT	Avg.
SOTA	Llama2-ISEAR	.99	.92	.49	.80
	Llama2-TEC	.98	.97	.55	.83
	Llama2-AT	.96	.94	.60	.83
	Mistral-ISEAR	.99	.95	.46	.80
	Mistral-TEC	.99	.97	.57	.84
	Mistral-AT	.98	.95	.63	.85
	GPT-3.5-ISEAR	.99	.90	.83	.90
	GPT-3.5-TEC	.94	.97	.70	.87
	GPT-3.5-AT	.97	.91	.88	.92
MOPO	GPT-3.5-All	.99	.97	.96	.97
	Llama2-All	.99	.97	.94	.96
	Mistral-All	.99	.97	.69	.88

Summary

- We automatically optimize prompts for better task performance.
- Multi-objective optimization across various (related) objectives: leads to better performance.

Outline

Zero-Shot (and few-shot) Predictions with Instruction-tuned Large Language Models

Instruction	Labels	Instance
Classify the text. Labels:	{positive, negative}.	``Vegetarian frikandel makes me happy.''

All these elements can be optimized or manually tuned for better predictions!

Prompt optimization

Optimize Label Verbalization

???

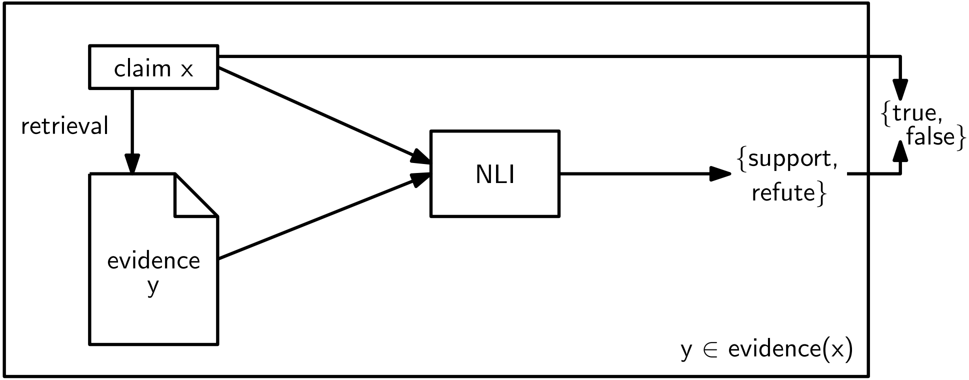
See for instance: Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, Sameer Singh (2020): AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP.

Self-Adaptive Paraphrasing and Preference Learning for Improved Claim Verifiability

Amelie Wührl[1,2] and **Roman Klinger**[2]
[1]University of Stuttgart, Germany [2]University of Bamberg, Germany
`firstname.lastname@uni-bamberg.de`

https://arxiv.org/abs/2412.11653

Evidence-based Fact-Checking with Black-Box NLI Model

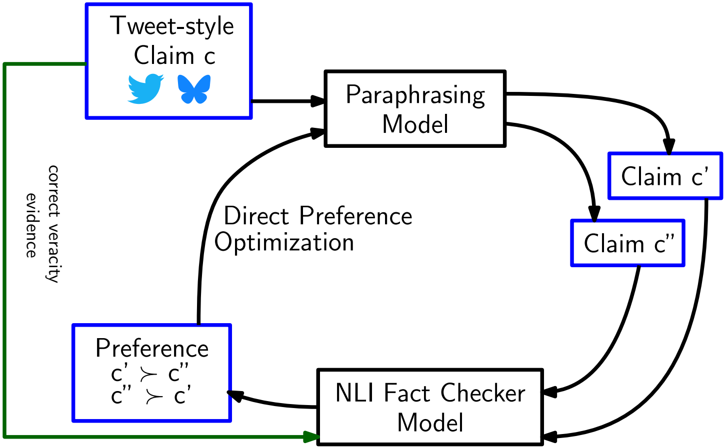


Challenge: Style mismatch between claim and evidence, but NLI models are typically trained on high-quality data sets.

Question and Approach

- Can we develop a paraphrasing model that reformulates claims such that the NLI model works better?
- Can we do so without “gold” claim reformulations?

Method: Direct Preference Optimization



R. Rafailov et al. (2023). "Direct Preference Optimization: Your Language Model is Secretly a Reward Model". In: NeurIPS

Experimental Setting

- Data: HealthVer Corpus of claims with evidences
- Prompt Llama-3-8B-Instruct to convert Claim to Tweet (with persona-based prompting)
- Fact checker: mDeBERTa (multilingual NLI, based on RoBERTa)
- Paraphraser: Prompt Llama-3-8B-Instruct with paraphrasing prompt
  ("Your task is to extract the checkworthy claim...)

M. Sarrouti et al. (2021). "Evidence-based Fact-Checking of Health-related Claims". In: Findings EMNLP

Results

- Original claim from data set: .47 weighted F1
- Tweet version of claim: .34 F1
- Zero-shot claim extraction without DPO: 0.40 F1
- DPO iterations: .43 F1
- ...not a huge improvement, but we see that paraphrasing the input for the NLI model does improve the result.

Example

Claim: Drinking boiled garlic water will cure COVID-19.

Evidence: In conclusion, Allium sativum may be an acceptable preventive measure against COVID-19 infection to boost immune system cells and to repress the production and secretion of proinflammatory cytokines as well as an adipose tissue derived hormone leptin having the proinflammatory nature

Tweet: Just saw someone claiming that sipping on boiled garlic water is the magic cure for COVID-19 Anyone else heard this one? Not buying it, folks! As a student, I've got my facts straight and there's no scientific evidence to back this up. Stay safe, stay informed, and let's not spread misinformation! #COVID19 #StaySafe #FactCheck

DPO_1: There is no evidence to support the claim that sipping on boiled garlic water is the magic cure for COVID19. It is likely a form of misinformation that can be harmful if believed and spread.

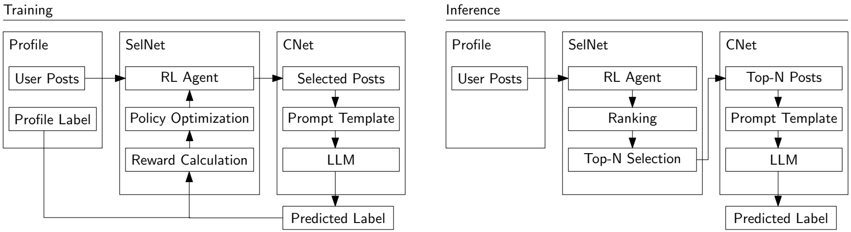
DPO_2: Sipping on boiled garlic water is a magic cure for COVID-19

Similar Setup: Reinforcement Learning to Filter Posts for Social Media User Profiling

Prompt-based Personality Profiling:
Reinforcement Learning for Relevance Filtering

Jan Hofmann[1], Cornelia Sindermann[2], and Roman Klinger[3]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Computational Digital Psychology, Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
[3]Fundamentals of Natural Language Processing, University of Bamberg, Germany
jan.hofmann@ims.uni-stuttgart.de; roman.klinger@uni-bamberg.de
cornelia.sindermann@iris.uni-stuttgart.de

Similar Setup: Reinforcement Learning to Filter Posts for Social Media User Profiling



Outline

Take Home

Making use of LLM for NLP tasks can be performed in various ways:

- Experiment with prompts manually.
  - $\Rightarrow$ Reasonable if task is not too sophisticated and can be introspected by a human
- Train parameters of the model
  - $\Rightarrow$ Reasonable if data, model weights and compute resources are available
- Optimize the prompt
  - $\Rightarrow$ Parameter-efficient optimization method, always doable, even without direct access to model
- Optimize the input data representation
  - Reasonable if there is a reason to believe that the input should not be used as is.

Thank you for
your attention.
Questions? Remarks?

?

Thanks to

- Yarik Menchaca Resendiz
- Amelie Wührl
- Lynn Greschner
- Jan Hofmann
- Cornelia Sindermann
- Generally: All of BamNLP (Bamberg) and IMS (Stuttgart)

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

University of Bamberg

Adapting Language Models for the Analysis of Real World Textual Data

Train the model, change the prompt, or adapt the data?

https://www.romanklinger.de/talks/klinger-koeln-2025.pdf

Cologne, May 22, 2025

Roman Klinger
roman.klinger@uni-bamberg.de

romanklinger.de romanklinger
https://www.bamberg.de/nlproc/