Automatic Term Extraction based on GenAI

Terminology Agent System

Bachelor's Thesis Presentation in Computer Science

Jonas Ludwig Gerg

24.03.2025 - 24.07.2025

Chair of Computer Systems
https://dse.in.tum.de/

Outline

Motivation

- Deutsche Bahn is facing a strong demand for qualified workers[1]
- Focus on recruitment from abroad[2]
- Language barriers due to domain-specific terminology
- Provision of explanations in texts and conversations
- Challenge: **How can we efficiently and reliably detect and extract terms and phrases in texts?**

[1] [1] “DB Job Portal aktuelle Anzahl an gesuchten Vollzeitstellen.” Accessed: Mar. 12, 2025. [Online]. Available: https://db.jobs/de-de/Suche

[2] [2] “Deutsche Bahn rekrutiert Auszubildende im Ausland.” Accessed: Jul. 02, 2025. [Online]. Available: https://www.sueddeutsche.de/wirtschaft/bahn-azubis-migration-1.5672979

Outline

Automatic Term Extraction (ATE)

The process of identification and **extraction** of **terminology** specific to a particular **domain** within a **document** using an **automated process**.[1]

[1][3] G. M. Di Nunzio, S. Marchesin, and G. Silvello, "A Systematic Review of Automatic Term Extraction: What Happened in 2022?," *Digital Scholarship in the Humanities*, vol. 38, no. Supplement_1, pp. i41–i47, Jun. 2023, doi: 10.1093/llc/fqad030.

[2]Example using nltk (https://www.nltk.org) and the universal tag set

Automatic Term Extraction (ATE)

The process of identification and **extraction** of **terminology** specific to a particular **domain** within a **document** using an **automated process**.[1]

- Linguistic: *"analyze linguistic properties of language"* [2]

```
[('Auf', 'NOUN'), ('dieser', 'NOUN'), ('Strecke', 'NOUN'), ('musst', 'NOUN'), ('du', 'NOUN'), ('auf',
'NOUN'), ('Sicht', 'NOUN'), ('fahren', 'NOUN'), ('!', '.')]
```

- Dictionary based: *"match with known dictionary entries"*
- Statistical: *"extract based on token frequency"*
- Large Language Models: *"extract using pretrained models"*

[1] [3] G. M. Di Nunzio, S. Marchesin, and G. Silvello, "A Systematic Review of Automatic Term Extraction: What Happened in 2022?," *Digital Scholarship in the Humanities*, vol. 38, no. Supplement_1, pp. i41–i47, Jun. 2023, doi: 10.1093/llc/fqad030.

[2] Example using nltk (https://www.nltk.org) and the universal tag set

Terms and Phrases in Texts

“Auf dieser Strecke musst du **auf Sicht fahren!**”

“**Fahrt auf Sicht** bis Kilometer 103.”

“Ich **fahre** hier seit längerem **auf Sicht**.”

“Ich muss **auf Sicht** für 3km **fahren**.”

Terms and Phrases in Texts

“Auf dieser Strecke musst du **auf Sicht fahren!**”

“**Fahrt auf Sicht** bis Kilometer 103.”

“Ich **fahre** hier seit längerem **auf Sicht**.”

“Ich muss **auf Sicht** für 3km **fahren**.”

→ **Fahrt auf Sicht**

Terms and Phrases in Texts

“Auf dieser Strecke musst du **auf Sicht fahren!**”

“**Fahrt auf Sicht** bis Kilometer 103.”

“Ich **fahre** hier seit längerem **auf Sicht**.”

“Ich muss **auf Sicht** für 3km **fahren**.”

→ **Fahrt auf Sicht**

→ **Auf Sicht fahren**

Terms occur in different variations in different languages

Challenges with ATE[1]

- Variability of Terms and Polysemy
- Domain Dependency
- Cross-Language Compatibility
  $\longrightarrow$ *different linguistic features*
- Scalability issues
  $\longrightarrow$ *especially dictionary based methods*

[1] [4] R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics 2nd Edition*, 2nd ed. Oxford University Press, 2014. doi: 10.1093/oxfordhb/9780199573691.001.0001.

Outline

Requirements and Design Goals

- Extensibility for new models, algorithms and data sources
- Reliable and safe definition generation
- Performant extraction and generation

Requirements and Design Goals

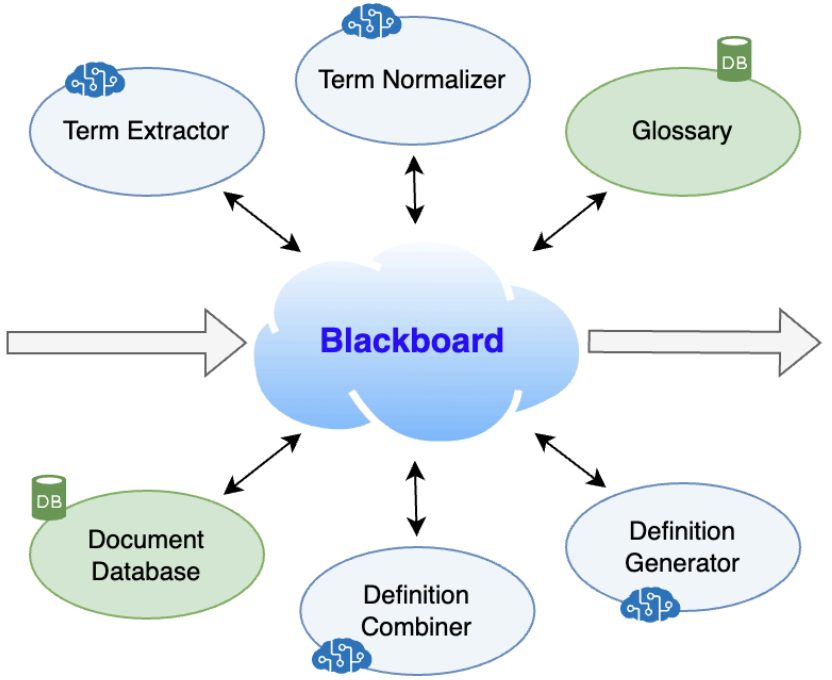
- Extensibility for new models, algorithms and data sources
- Reliable and safe definition generation
- Performant extraction and generation

→ Main priority on reliability and safety

Outline

Design Overview

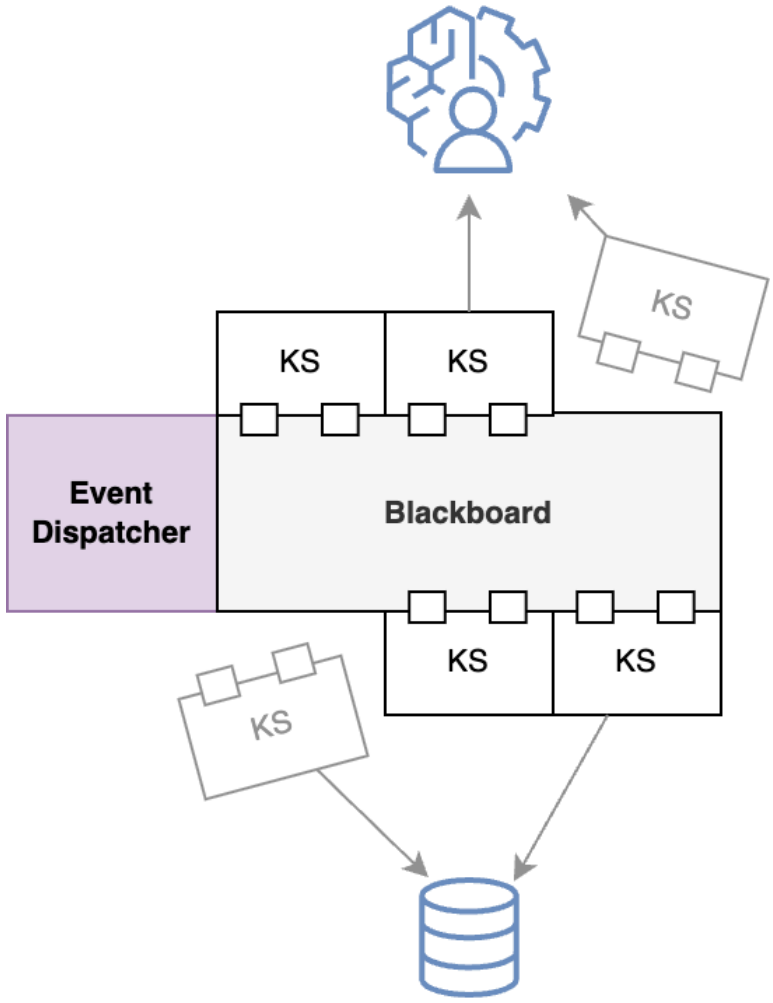
"Ich fahre hier seit längerem auf Sicht."




```
{
  "terms": ["auf Sicht", "Fahrt auf Sicht"],
  "sources": [
    ...
  ]
}
```

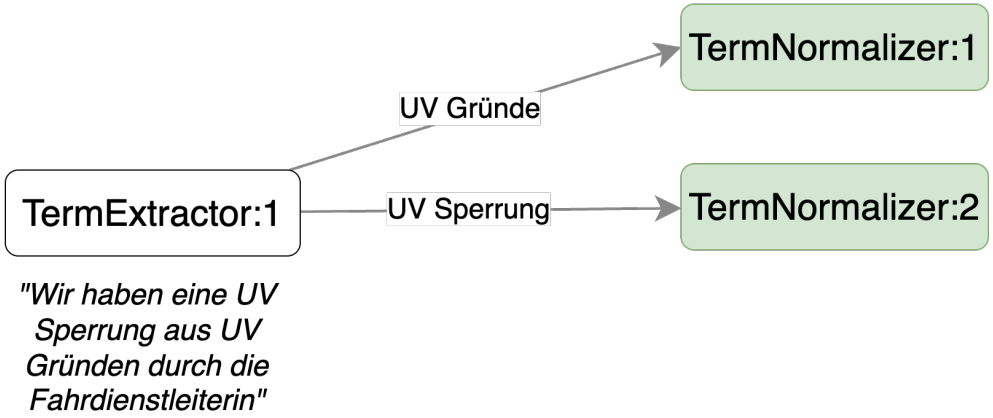
Design Details

- Event-driven blackboard pattern architecture[1]
- Knowledge Sources (KS) …
  - … operate on shared blackboard
  - … react to and publish new events
  - … run independently
  - … employ **LLMs** to solve linguistic tasks
  - … optionally access external systems

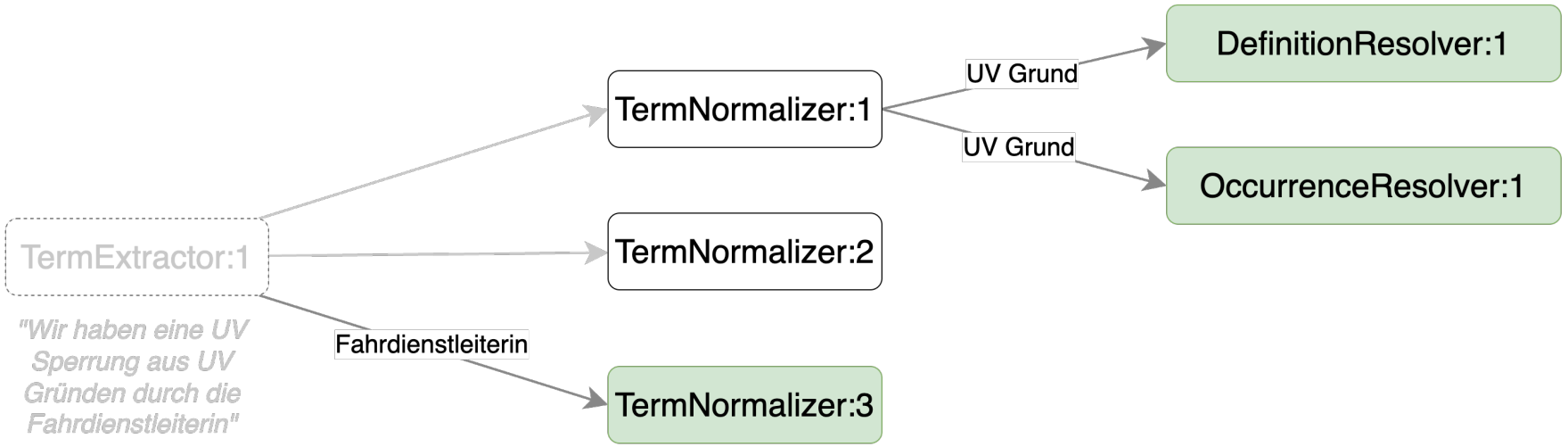


[1] [5] P. Lalanda, “Two Complementary Patterns to Build Multi-Expert Systems.”

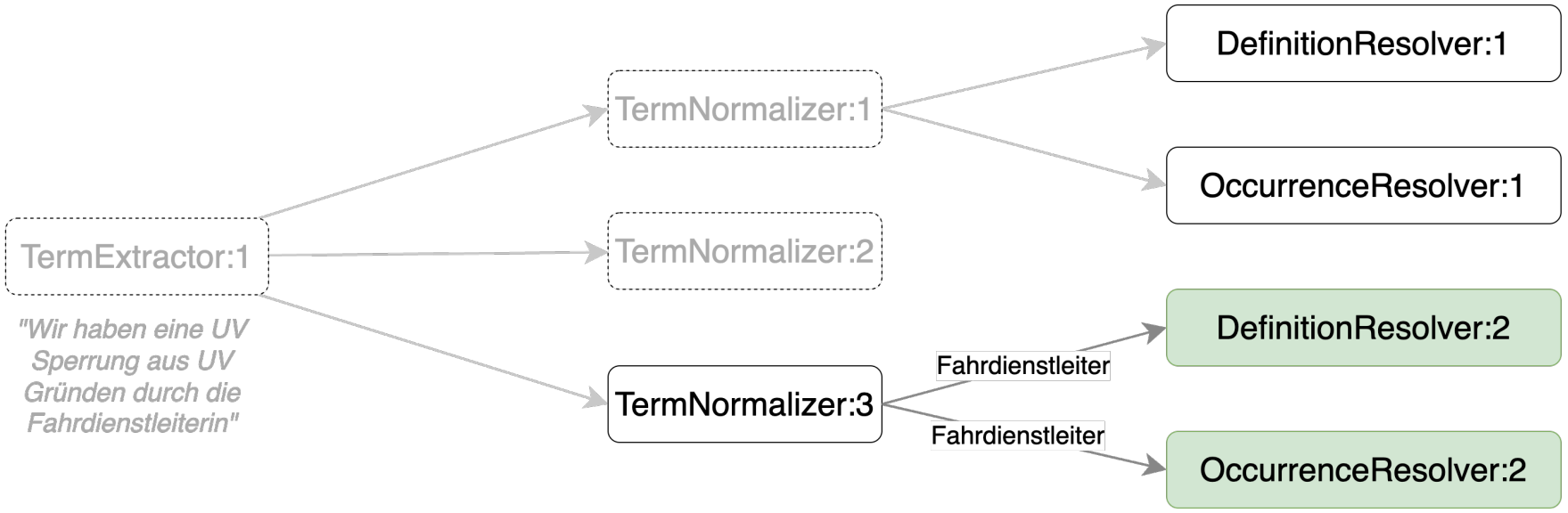
Example Event Flow



Example Event Flow

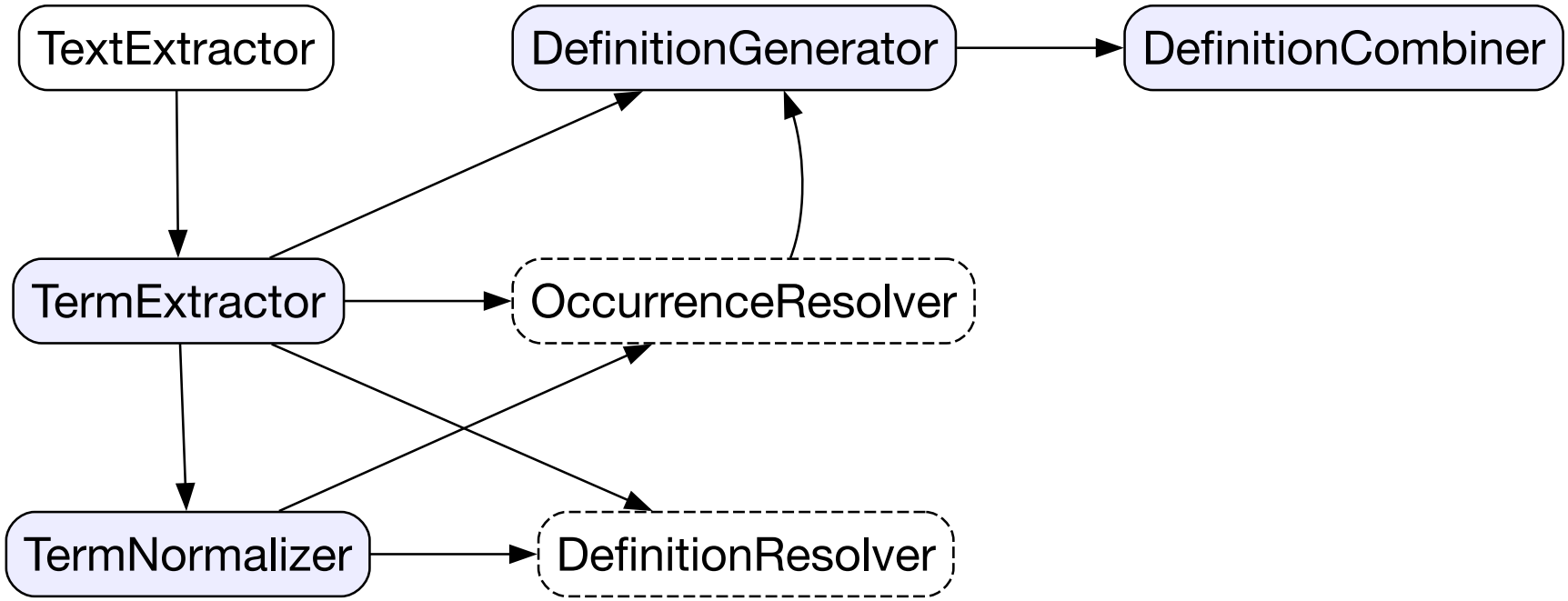


Example Event Flow



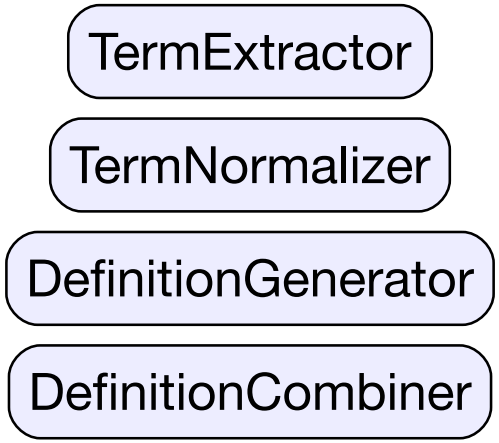
Outline

Knowledge Sources



LLM based Knowledge Sources

- Use OpenAIs `gpt-4o-mini` model
- Different Prompt Engineering techniques:[1]
  - Zero-Shot
  - Few-Shot
  - Prompt-Chaining[2]
- `temperature=0` ⟶ reduce randomness[3]
- Implementation for DB uses internal BahnGPT



[1] [6] *Prompt Engineering for LLMs*. Accessed: Jun. 11, 2025. [Online]. Available: https://learning.oreilly.com/library/view/prompt-engineering-for/9781098156145/

[2] [7] Anthropic, "Chain Complex Prompts for Stronger Performance." Accessed: Jul. 04, 2025. [Online]. Available: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/chain-prompts

[3] [8] M. Renze, "The Effect of Sampling Temperature on Problem Solving in Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 7346–7356. doi: 10.18653/v1/2024.findings-emnlp.432.

Outline

Testing LLM Applications

```js
// probabilities in iteration 1
{'FALSE': 0.6224593312018545, 'TRUE': 0.37754066879814 54}
// probabilities in iteration 2
{'FALSE': 0.9149009474519222, 'TRUE': 0.08509905254807776}
// probabilities in iteration 3
{'TRUE': 0.679178699175393, 'FALSE': 0.32082130082460697}
// probabilities in iteration 4
{'FALSE': 0.9399133527714579, 'TRUE': 0.060086647228542005}
// probabilities in iteration 5
{'TRUE': 0.9241418131886822, 'FALSE': 0.0758581868113179}
```

→ Multiple iterations of tests were not consistent, even with deterministic LLM settings

Challenges with Testing of LLM Applications

1. **Non-Determinism** and Inconsistency of LLM output
   → *different output for every execution*
2. **Semantic comparison** between oracle and observation
   → *colloquial **“420er”** ↔ official **“420”***
3. **Increased cost** when allowing reasoning
   → *however: testing was successful afterwards*

Evaluation

- Limited test set of 21 crafted, short test sentences

Outline

Summary

Traditional ATE methods not properly applicable

- limited cross-language support
- domain-specificity
- fine-tuning and training required

TAS:

- Flexible integration of new technologies or data sources
- Out-of-the-box models
- Potentially high recall rate

⚠️ Further research required in determinism for testing LLM applications and the performance of TAS

⏩ Future Research

- Performance improvements and scalability of TAS
  - Fine-tuning of baseline models
  - Distributed knowledge sources
- Cross-language and cross-domain compatibility
- ChatBot functionality for TAS
- Testing of LLM applications (*“how to deal with non-determinism?”*)

Sources

[1] “DB Job Portal aktuelle Anzahl an gesuchten Vollzeitstellen.” Accessed: Mar. 12, 2025. [Online]. Available: https://db.jobs/de-de/Suche

[2] “Deutsche Bahn rekrutiert Auszubildende im Ausland.” Accessed: Jul. 02, 2025. [Online]. Available: https://www.sueddeutsche.de/wirtschaft/bahn-azubis-migration-1.5672979

[3] G. M. Di Nunzio, S. Marchesin, and G. Silvello, “A Systematic Review of Automatic Term Extraction: What Happened in 2022?,” *Digital Scholarship in the Humanities*, vol. 38, no. Supplement_1, pp. i41–i47, Jun. 2023, doi: 10.1093/llc/fqad030.

[4] R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics 2nd Edition*, 2nd ed. Oxford University Press, 2014. doi: 10.1093/oxfordhb/9780199573691.001.0001.

[5] P. Lalanda, “Two Complementary Patterns to Build Multi-Expert Systems.”

[6] *Prompt Engineering for LLMs*. Accessed: Jun. 11, 2025. [Online]. Available: https://learning.oreilly.com/library/view/prompt-engineering-for/9781098156145/

[7] Anthropic, “Chain Complex Prompts for Stronger Performance.” Accessed: Jul. 04, 2025. [Online]. Available: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/chain-prompts

[8] M. Renze, “The Effect of Sampling Temperature on Problem Solving in Large Language Models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 7346–7356. doi: 10.18653/v1/2024.findings-emnlp.432.