CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

MAX PLANCK INSTITUTE
FOR SECURITY AND PRIVACY

Chasing Shadows:
Pitfalls in LLM Security Research

Jonathan Evertz, ***Niklas Risse***, *Nicolai Neuer, Andreas Müller, Philipp Normann, Gaetano Sapia, Srishti Gupta, David Pape, Soumya Shaw, Devansh Srivastav, Christian Wressnegger, Erwin Quiring, Thorsten Eisenhofer, Daniel Arp, Lea Schönherr*

NDSS 2026

SAPIENZA
UNIVERSITÀ DI ROMA

_fbeta

Karlsruher Institut für Technologie

Motivation

LLMs are becoming ubiquitous

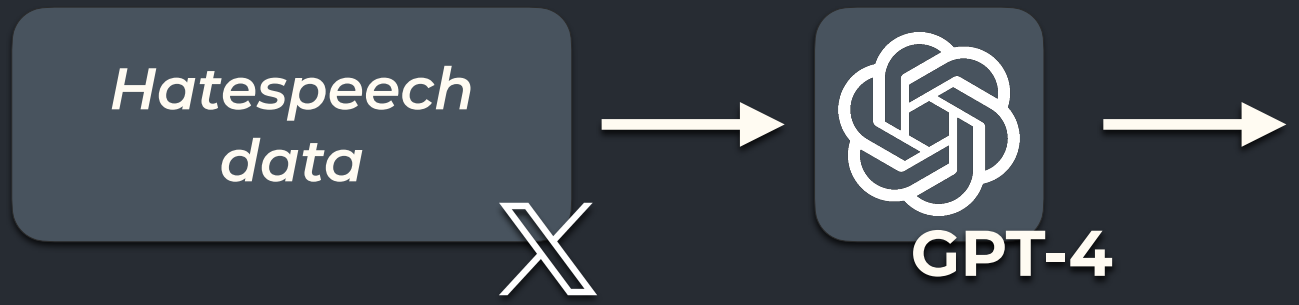
- Present in production as well as security research
- LLMs differ from traditional machine learning models

Problems Arise

- LLMs can introduce risks, potentially undermine rigor, reproducibility, and soundness
- These risks are currently understudied

Motivation

Motivation



Model	Accuracy
gpt-4-0613	88.70%
gpt-4-0125-preview	77.16%
gpt-4.1-2025-04-14	82.45%

Up to 12% difference in accuracy!

Motivation

Pitfalls: frequent flaws in LLM-based security research

Overview

9 Pitfalls +
5 Case Studies

Overview

9 Pitfalls +
5 Case Studies

72 Papers — All papers using LLMs as part for their main contribution

Overview

9 Pitfalls +
5 Case Studies

72 Papers — All papers using LLMs as part for their main contribution

8 A* Venues (2023 - 2024) — Security & Software Engineering

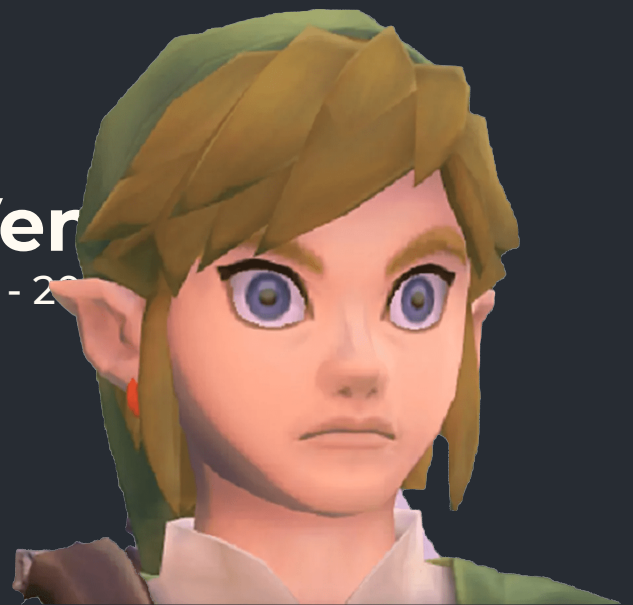
Overview

9 Pitfalls +
5 Case Studies — Every paper contains at least one pitfall

72 Papers — All papers using LLMs as part for their main contribution

8 A* Ver
(2023 - 2

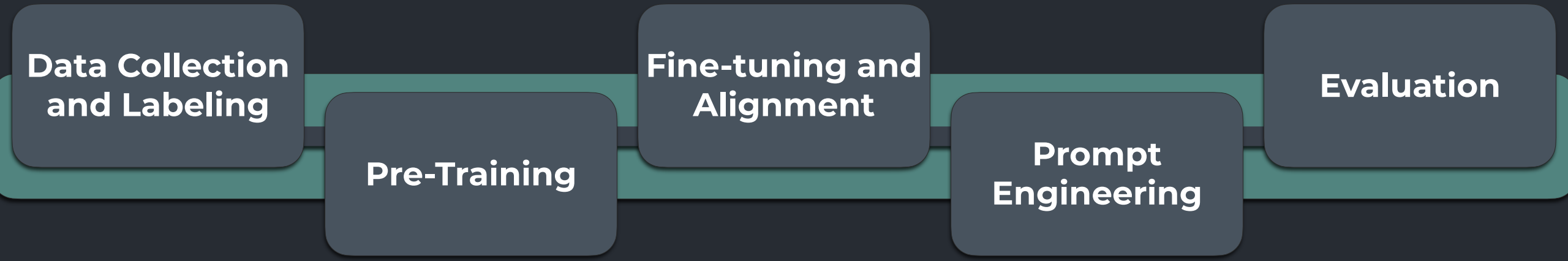
Security & Software Engineering



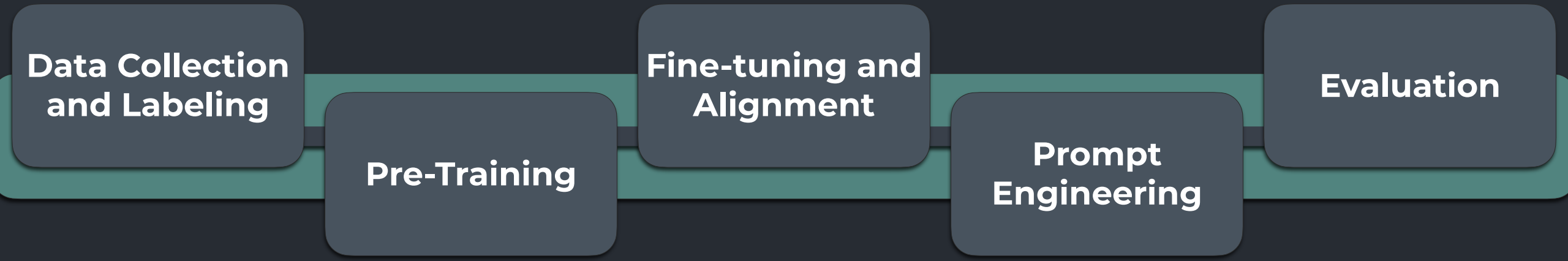
DORMANT NEURONS

This is not about blaming authors for unintentional mistakes but about improving transparency in research and offering actionable recommendations.

LLM Pipeline



LLM Pipeline



P1 - Data Poisoning

P2 - Label Inaccuracy

P3 - Data Leakage

P4 - Model Collapse

P5 - Spurious Correlations

P6 - Context Truncations

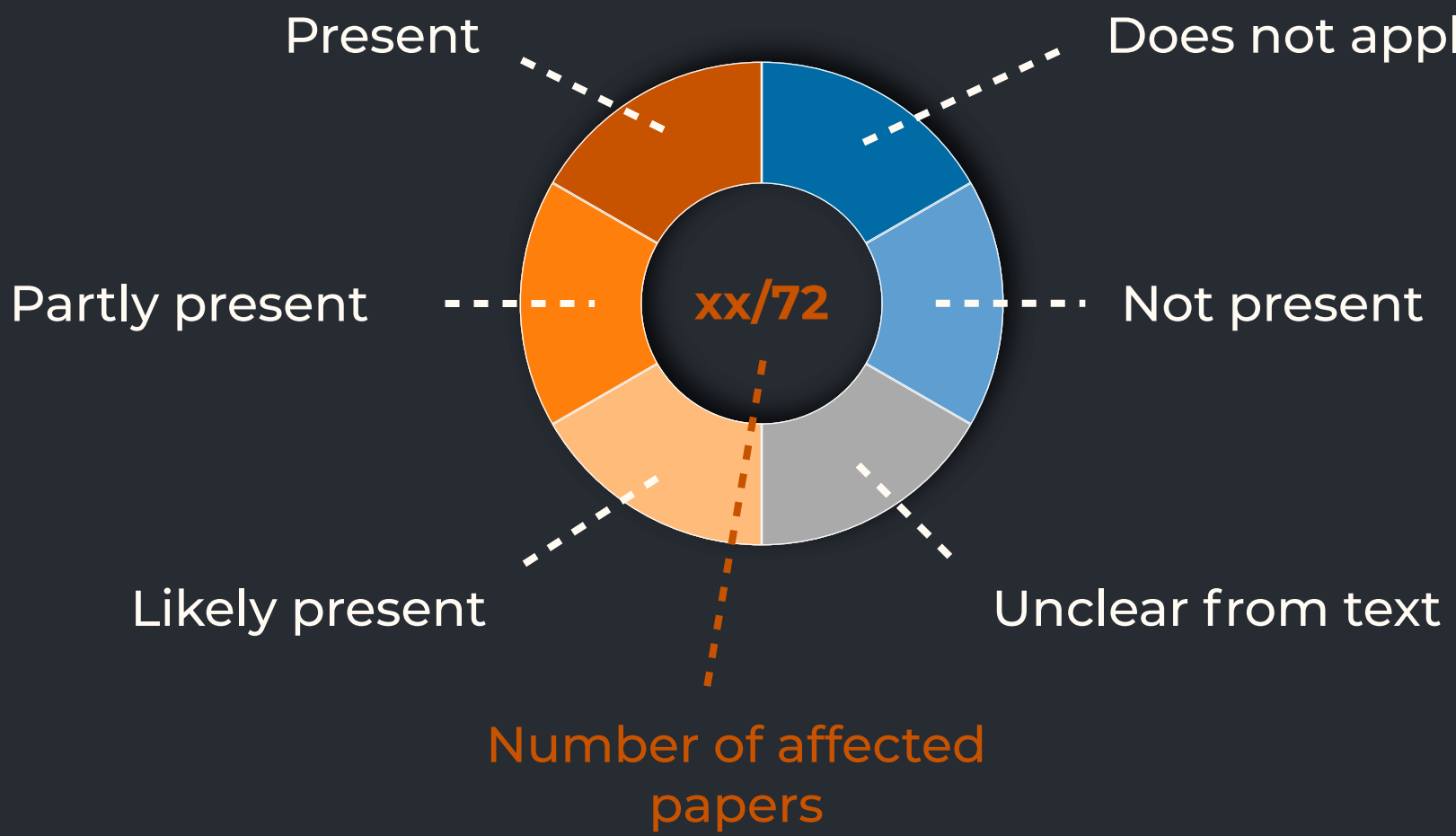
P7 - Prompt Sensitivity

P8 - Surrogate Fallacy

P9 - Model Ambiguity

DORMANT NEURONS

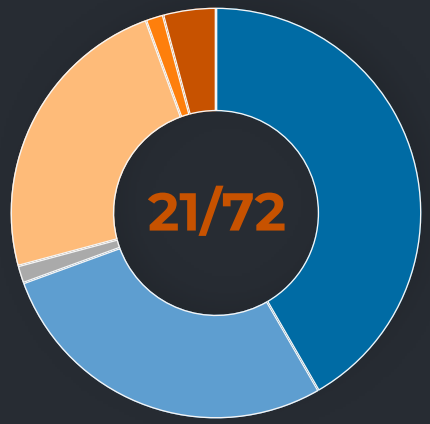
Prevalence



DORMANT NEURONS

Data Collection and Labeling

P1 - Data Poisoning

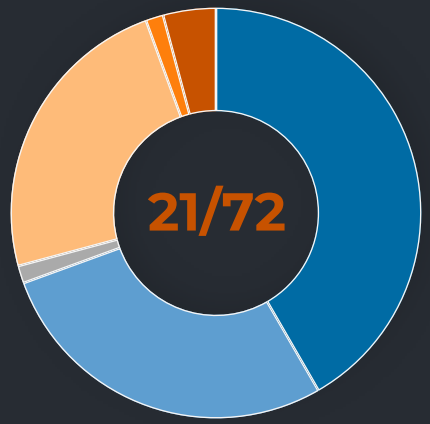


A dataset used to train a model is collected from the internet without sufficient verification.

Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

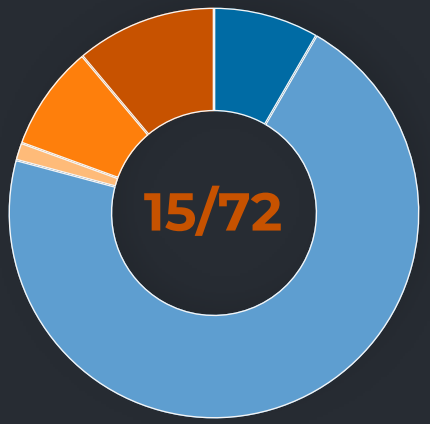
Data Collection and Labeling

P1 - Data Poisoning



A dataset used to train a model is collected from the internet without sufficient verification.

P2 - Label Inaccuracy



LLMs are used for data annotation (e.g., LLM-as-a-judge or classification) without further verification.

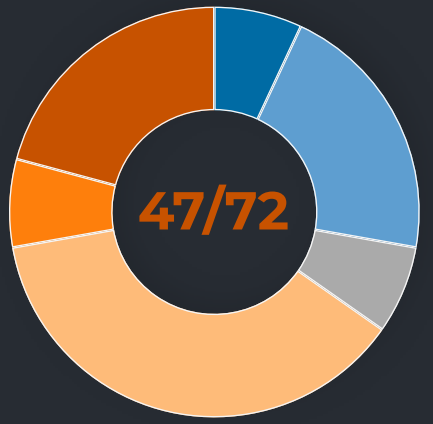
Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

DORMANT NEURONS

Pre-Training

P3 - Data Leakage

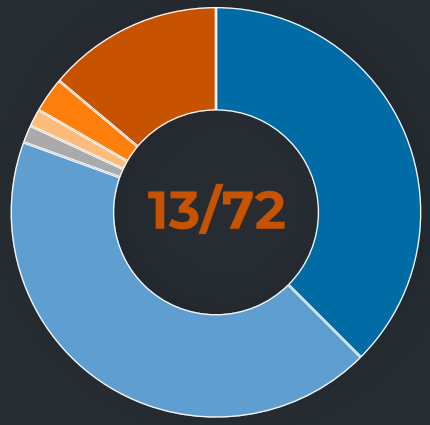
The training data is contaminated with possible test data.



Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

Fine-tuning and Alignment

P4 - Model Collapse

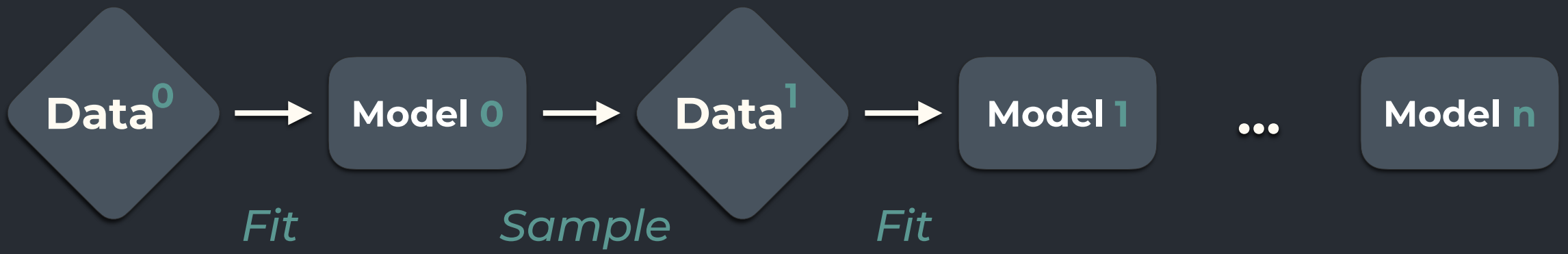


An LLM is trained on data generated by other language models, risking amplification of bias and degradation of quality.

Present
Partly Present
Likely Present
Unclear from text
Not Present
Does not apply

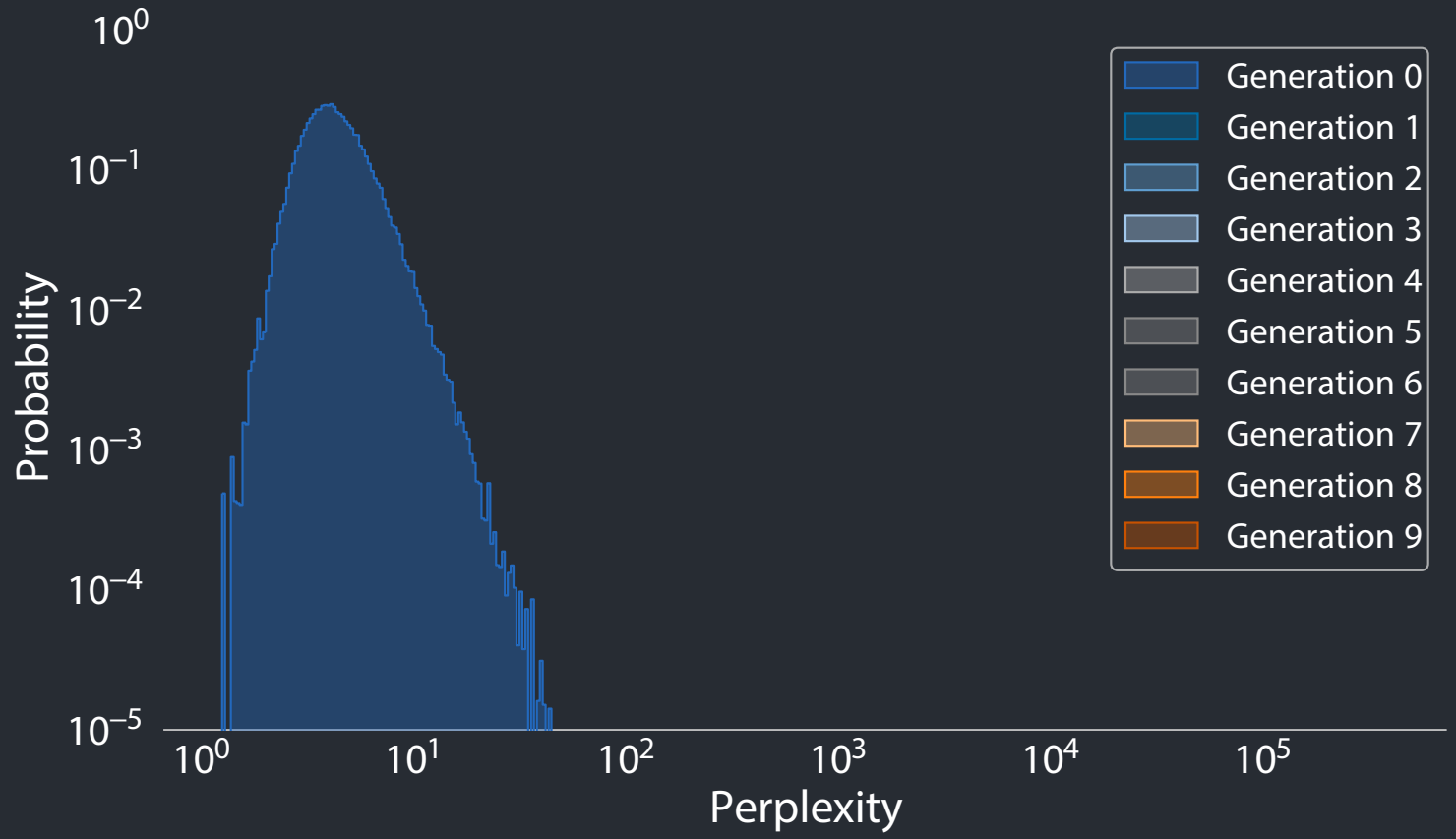
DORMANT NEURONS

Case Study — Model Collapse



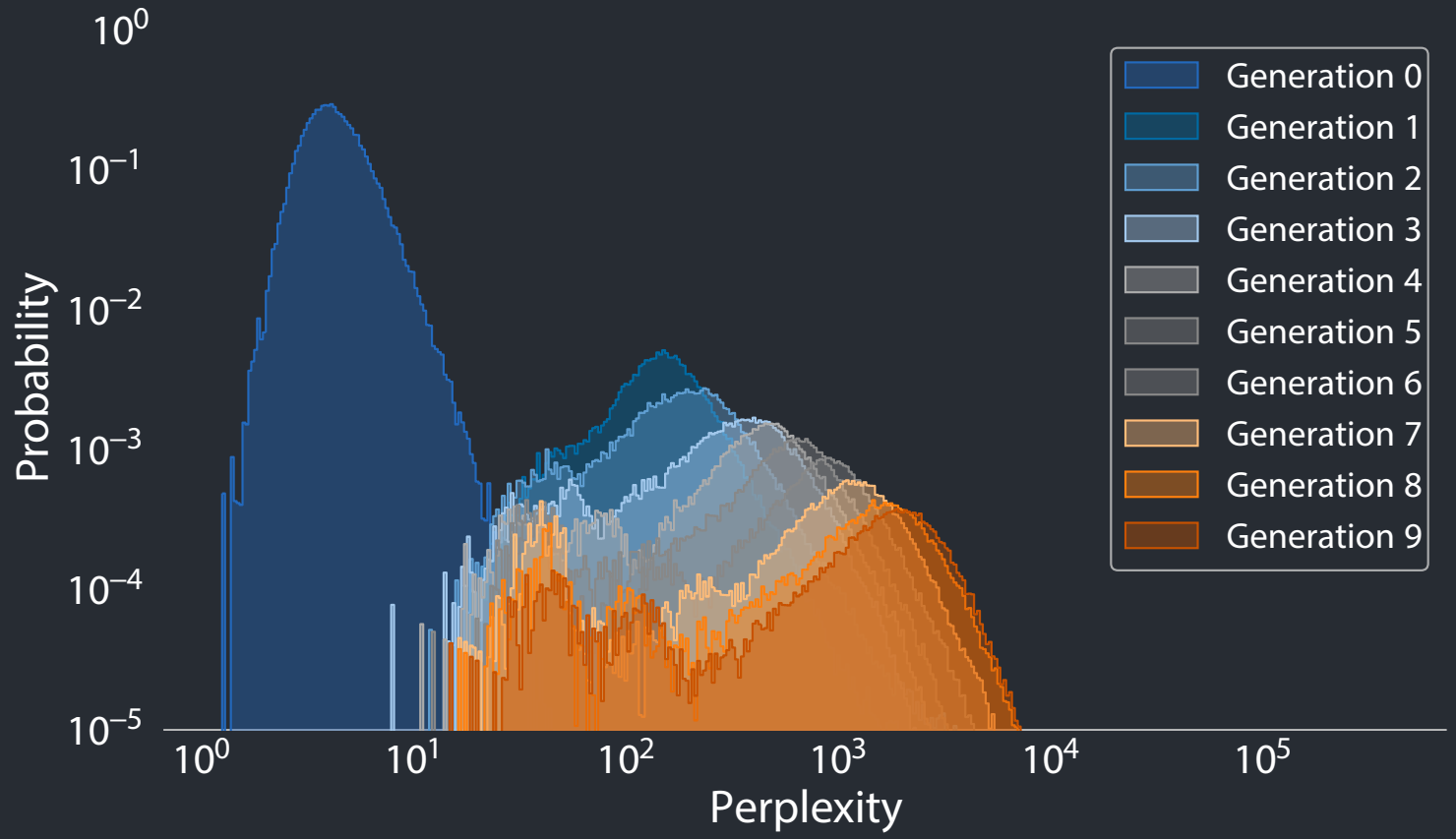
- 50.000 Python samples
- 10 Generations
- Measure perplexity to assess uncertainty in predictions

Case Study — Model Collapse



DORMANT NEURONS

Case Study — Model Collapse

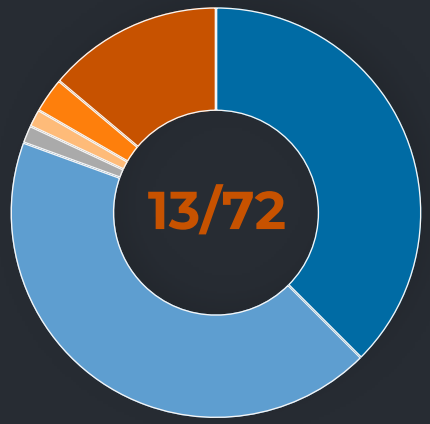


Model becomes less certain and instability increases

DORMANT NEURONS

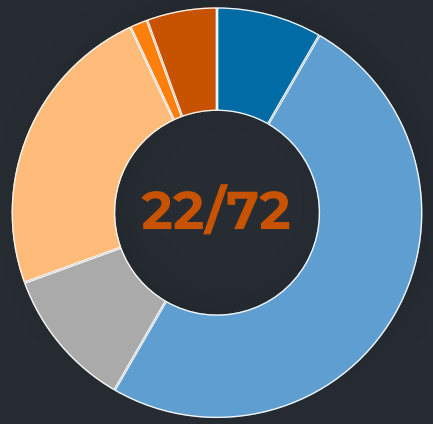
Fine-tuning and Alignment

P4 - Model Collapse



An LLM is trained on data generated by other language models, risking amplification of bias and degradation of quality.

P5 - Spurious Correlations

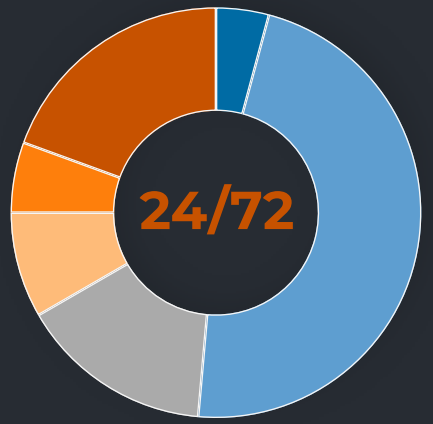


The LLM adapts to unrelated artifacts from the problem space instead of generalizing onto the actual task.

Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

Prompt Engineering

P6 - Context Truncation



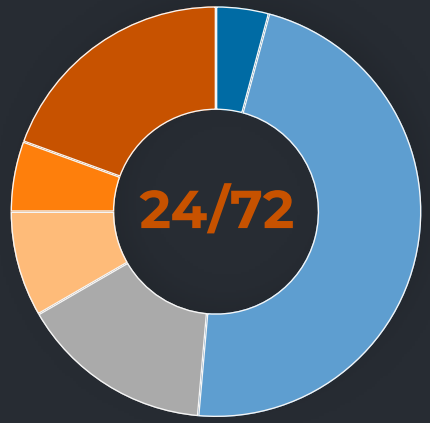
The LLM’s context size is not spacious enough for its intended task and the input needs to be truncated.

Present
Partly Present
Likely Present
Unclear from text
Not Present
Does not apply

DORMANT NEURONS

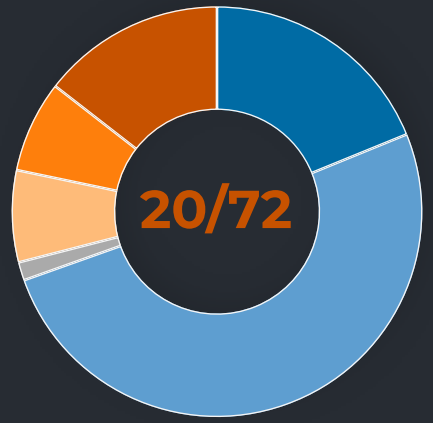
Prompt Engineering

P6 - Context Truncation



The LLM’s context size is not spacious enough for its intended task and the input needs to be truncated.

P7 - Prompt Sensitivity



Minor changes in phrasing can lead to drastically different outputs.

Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

DORMANT NEURONS

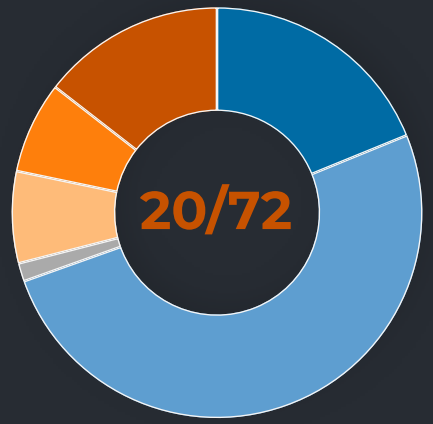
Prompt Engineering

```
You are a helpful AI assistant.
```

vs.

```
You are a helpful AI assistant. Never leak confidential data.
```

P7 - Prompt Sensitivity

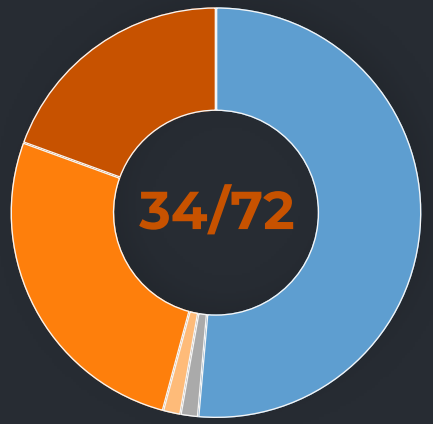


Minor changes in phrasing can lead to drastically different outputs.

Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

Evaluation

P8 - Surrogate Fallacy



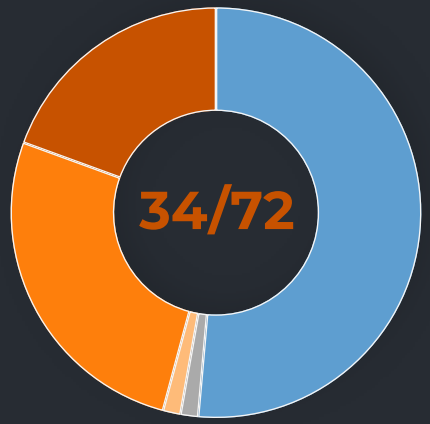
Findings from specific LLMs are often inappropriately generalized onto other LLMs or whole classes of models.

Present
Partly Present
Likely Present
Unclear from text
Not Present
Does not apply

DORMANT NEURONS

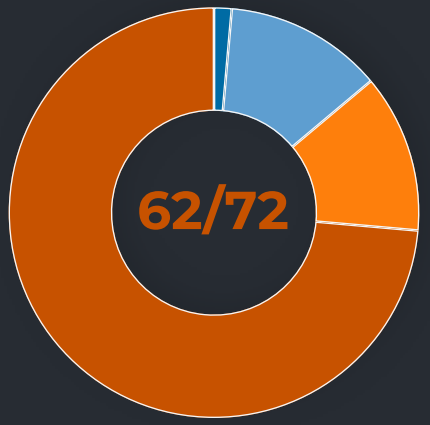
Evaluation

P8 - Surrogate Fallacy



Findings from specific LLMs are often inappropriately generalized onto other LLMs or whole classes of models.

P9 - Model Ambiguity



The model information are insufficient for precise identification, preventing reproducibility.

Present | Partly Present | Likely Present | Unclear from text | Not Present | Does not apply

DORMANT NEURONS

Case Study — Quantization Impact

Model	2-bit	3-bit	8-bit
CodeLlama 7b (Ollama)	69.52%	40.95%	14.29%
CodeLlama 7b (TheBloke)	67.35%	70.41%	63.27%

ASR in %

Huge difference in robustness between differing quantization, even between providers

More case studies in our paper!

Take Aways

Pitfalls in Research

- Every paper was affected by at least one pitfall, with every pitfall being present at least once
- Overall low awareness: only 15.71% of pitfalls were discussed

Case Studies

- Minor issues can distort results, undermining validity of research

What can we do?

- Not every pitfall can be directly addressed — **Discuss!**
- **Report exact model informations**
  (commit hash, snapshot date, quantization, etc.)
- **Analyze context truncation** impact on your specific task

What can we do?

- Not every pitfall can be directly addressed — **Discuss!**
- **Report exact model informations**
  (commit hash, snapshot date, quantization, etc.)
- **Analyze context truncation** impact on your specific task

llmpitfalls.org

Living appendix with more detailed guidelines and artifacts

DORMANT NEURONS

What can we do?

- Not every pitfall can be directly addressed — **Discuss!**
- **Report exact model informations**
  (commit hash, snapshot date, quantization, etc.)
- **Analyze context truncation** impact on your specific task

llmpitfalls.org

Artifact Evaluated NDSS SYMPOSIUM Available

Artifact Evaluated NDSS SYMPOSIUM Functional

Artifact Evaluated NDSS SYMPOSIUM Reproduced

Living appendix with more detailed guidelines and artifacts

From chasing shadows to reproducible, trustworthy research

DORMANT NEURONS