# Supporting Artifact Evaluation – LLMs to the Rescue?

**Jan Pennekamp**

**RWTH Aachen University**
**Communication and Distributed Systems**

**pennekamp@comsys.rwth-aachen.de**

Learning from Authoritative Security Experiment Results (LASER) Workshop
December 9, 2025 – ACSAC – Honolulu, HI, USA

# A Short Disclaimer

- **Expectation management**
  - This talk is not your typical “keynote”
  - A critical view rather than a stringent success story
    - Highly experimental results
    - Food for thought






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# A Short Disclaimer

- **Expectation management**
  - This talk is not your typical “keynote”
  - A critical view rather than a stringent success story
    - Highly experimental results
    - Food for thought




- **Meet David Heye—first author of this work**





Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# A Short Disclaimer

- **Expectation management**
  - This talk is not your typical “keynote”
  - A critical view rather than a stringent success story
    - Highly experimental results
    - Food for thought

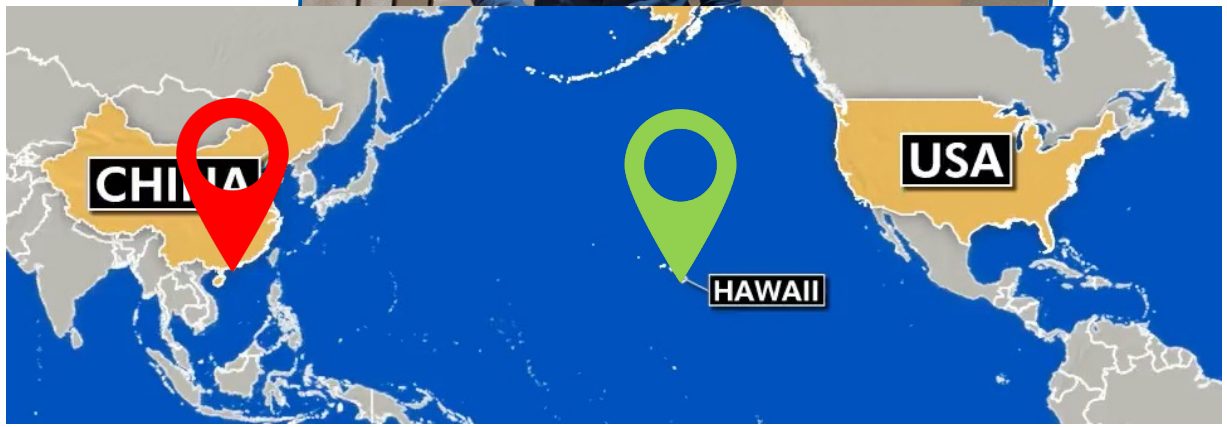


**Caution science in progress**

  - Earlier today, David presented this work at IEEE BigData 2025 in Macau, China

- **Meet David Heye—first author of this work**








Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# A Quick Tour down Memory Lane



- **This work is a true team effort**
  - Idea initially proposed by students as part of an interdisciplinary research-oriented teaching course
  - Thanks to David Heye, **Karl Kindermann**, **Robin Decker**, Johannes Lohmöller, Anastasiia Belova, Sandra Geisler, and Klaus Wehrle



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# A Quick Tour down Memory Lane



- **This work is a true team effort**
  - Idea initially proposed by students as part of an interdisciplinary research-oriented teaching course
  - Thanks to David Heye, **Karl Kindermann**, **Robin Decker**, Johannes Lohmöller, Anastasiia Belova, Sandra Geisler, and Klaus Wehrle

- **So, what brings me here?** My previous interaction with David Balenson and LASER:



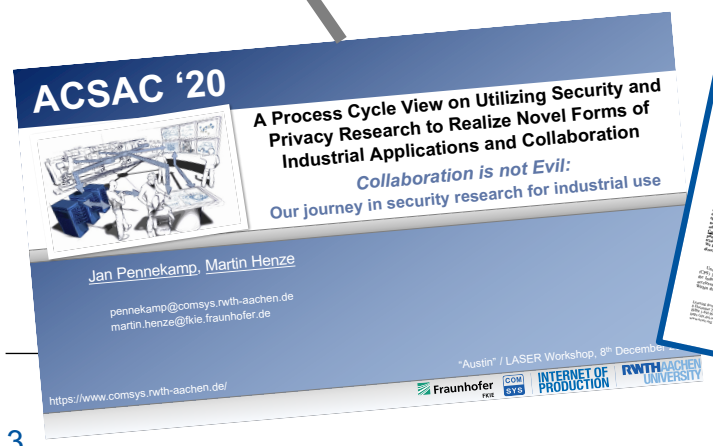
Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# A Quick Tour down Memory Lane

- **This work is a true team effort**
  - Idea initially proposed by students as part of an interdisciplinary research-oriented teaching course
  - Thanks to David Heye, **Karl Kindermann**, **Robin Decker**, Johannes Lohmöller, Anastasiia Belova, Sandra Geisler, and Klaus Wehrle

- **So, what brings me here?** My previous interaction with David Balenson and LASER:

**2020** **2021** **2022** **2023** **Today**
















Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Peer Review in Academia






Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Peer Review in Academia

Do Research






Write about it



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Peer Review in Academia

Do Research

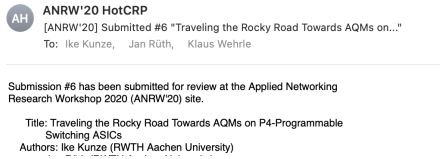





Write about it



Submit it for Publication



AH **ANRW'20 HotCRP**
[ANRW'20] Submitted #6 "Traveling the Rocky Road Towards AQMs on..."
To: Ike Kunze, Jan Rüth, Klaus Wehrle

Submission #6 has been submitted for review at the Applied Networking Research Workshop 2020 (ANRW'20) site.

Title: Traveling the Rocky Road Towards AQMs on P4-Programmable Switching ASICs
Authors: Ike Kunze (RWTH Aachen University)

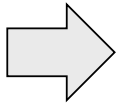


Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Peer Review in Academia

Do Research








Write about it



Submit it for Publication

AH **ANRW'20 HotCRP**
[ANRW'20] Submitted #6 "Traveling the Rocky Road Towards AQMs on..."
To: Ike Kunze, Jan Rüth, Klaus Wehrle

Submission #6 has been submitted for review at the Applied Networking Research Workshop 2020 (ANRW'20) site.

Title: Traveling the Rocky Road Towards AQMs on P4-Programmable Switching ASICs
Authors: Ike Kunze (RWTH Aachen University)

Peer Review



I **IM 2021**
Your IM 2021 paper 210992
To: kunze@comsys.rwth-aachen.de, Cc: Klaus Wehrle, Jan Rüth
Reply-To: im2021tpc@gmail.com

Dear Mr. Ike Kunze:

Your paper "ASIC + P4: A Strong Compound for AQM on High-Speed Networks?" (#210992) has been accepted to be presented at the main conference technical sessions of the 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM 2021). Congratulations!



AH **ANRW'20 HotCRP**
[ANRW'20] Rejected submission #6 "Traveling the Rocky Road Towa
To: Ike Kunze, Jan Rüth, Klaus Wehrle, Cc: Roland van Rijswijk
Reply-To: r.m.vanrijswijk@utwente.nl, Mirja Kuehlewind

Dear authors,

The Applied Networking Research Workshop 2020 (ANRW'20) program committee is sorry to inform you that your submission #6 was rejected for the workshop.

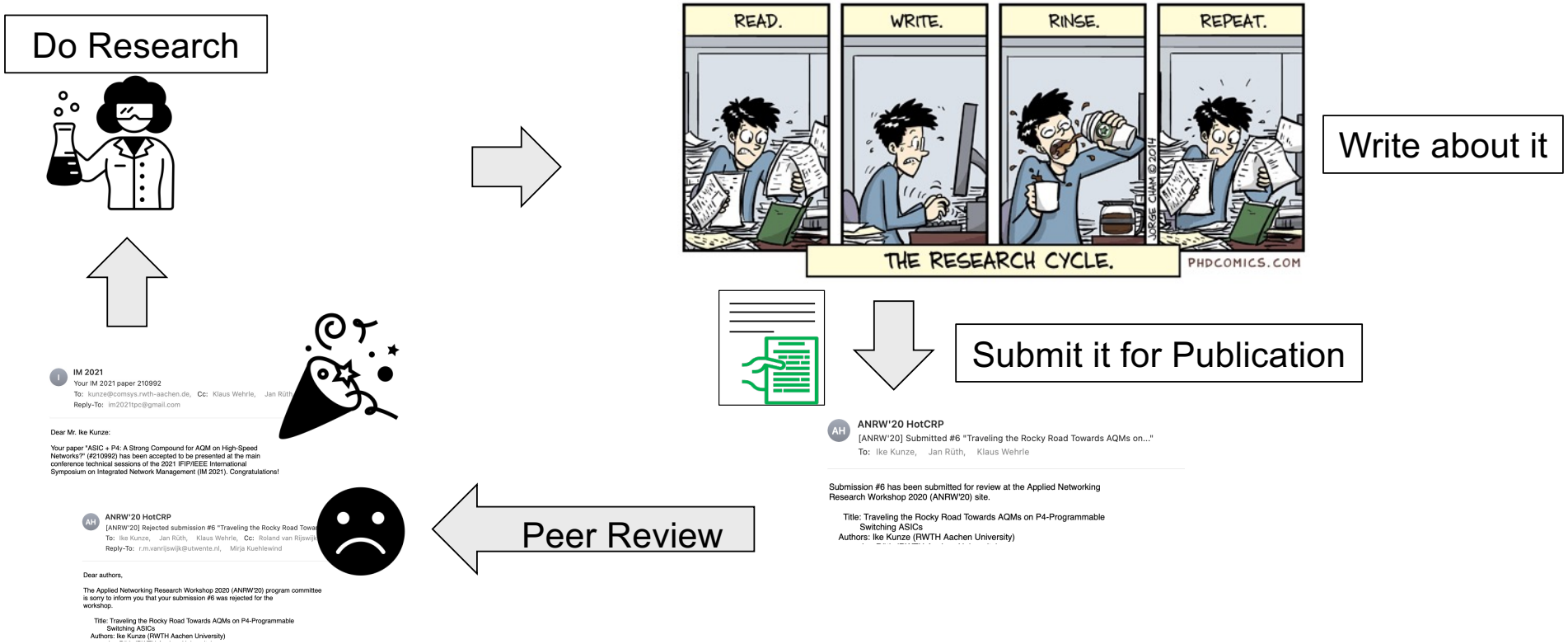
Title: Traveling the Rocky Road Towards AQMs on P4-Programmable Switching ASICs
Authors: Ike Kunze (RWTH Aachen University)



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

COMSYS | RWTH AACHEN UNIVERSITY

# Peer Review in Academia




- **Peer review is the best thing that we (currently) have**
  - Check suitability for publication (and soundness of the submitted work)
  - Help researchers to improve their work, including constructive criticism



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Challenge Growing Number of Submissions



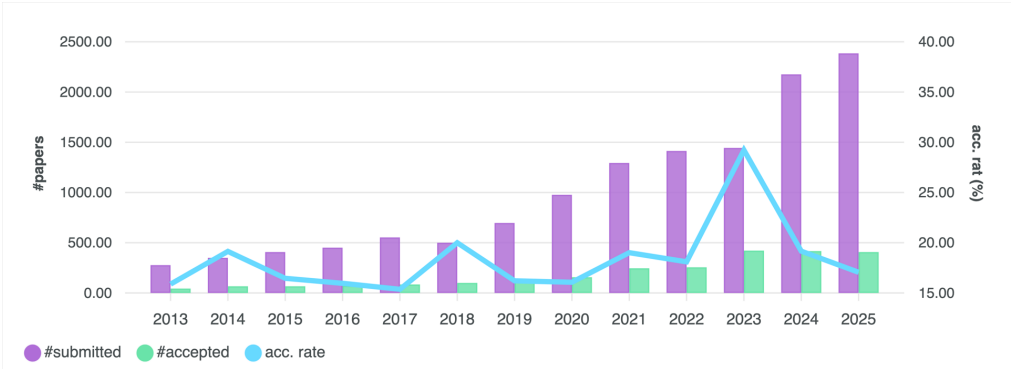
- **Are peer reviews still scaling?**
  - Growing number of researchers
    - Implies also an increase of submissions
  - Publish or perish culture



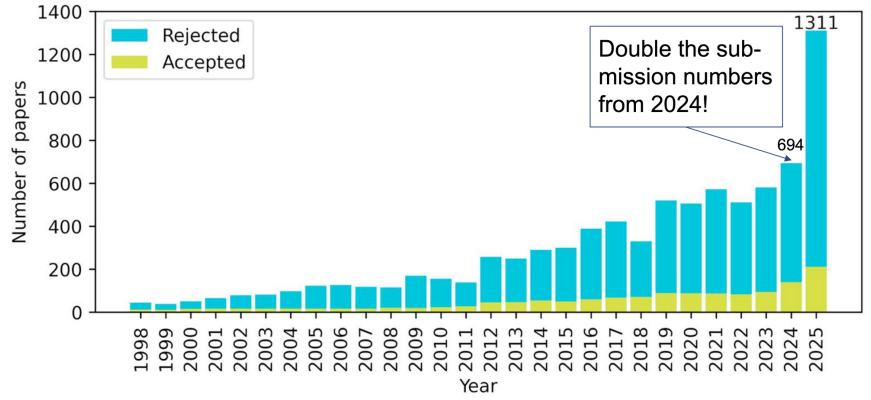
Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Challenge Growing Number of Submissions

- **Are peer reviews still scaling?**
  - Growing number of researchers
    - Implies also an increase of submissions
  - Publish or perish culture




Paper submissions at USENIX Security



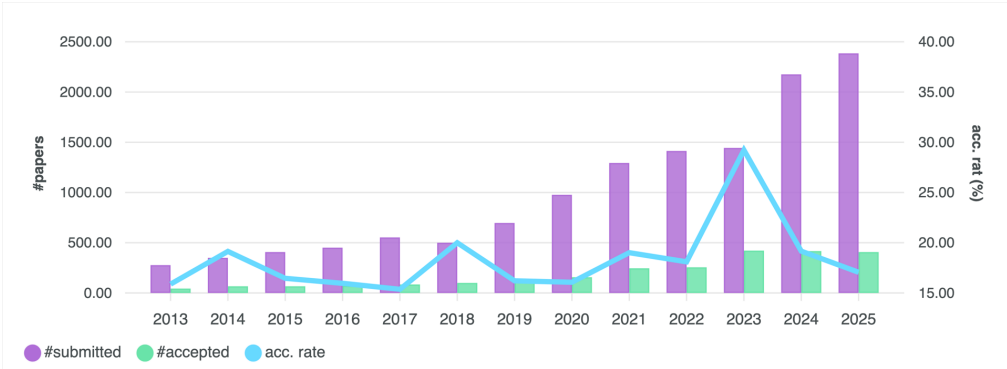

Paper submissions at NDSS



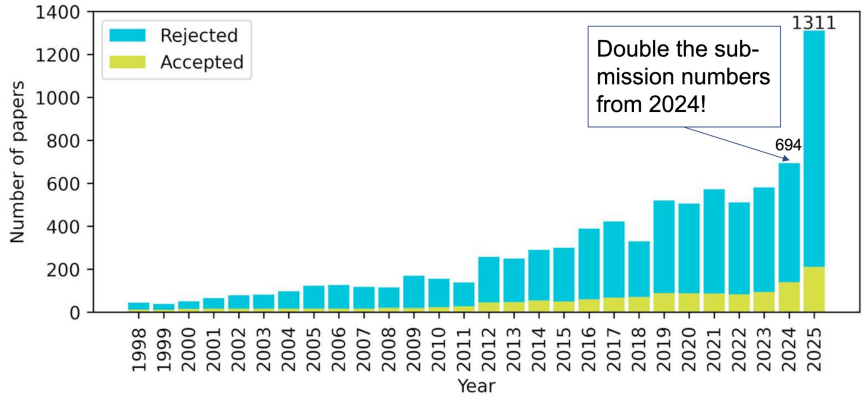
Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Challenge Growing Number of Submissions

- **Are peer reviews still scaling?**
  - Growing number of researchers
    - Implies also an increase of submissions
  - Publish or perish culture
  - PC sizes are growing in somewhat comparable pace
    - Can we still fulfill the promise of peer reviews, i.e., timely reviews by experts in the field?
      - Emergency PC members, subreviewing, etc.
      - Reviewing fatigue




Paper submissions at USENIX Security




Paper submissions at NDSS

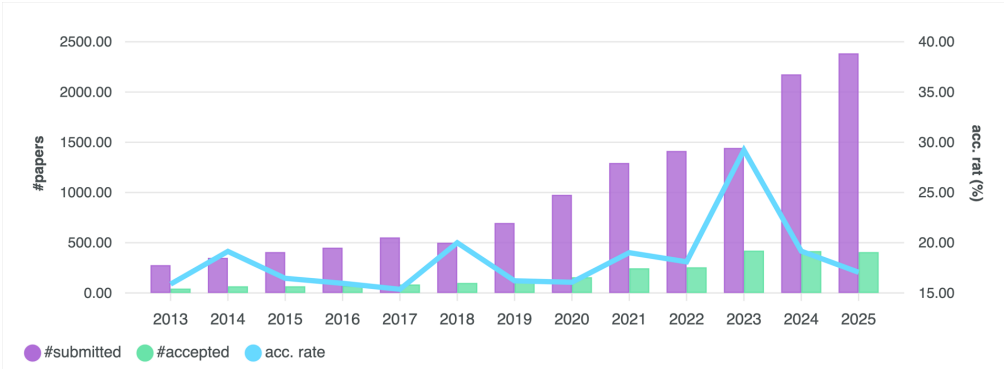


Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

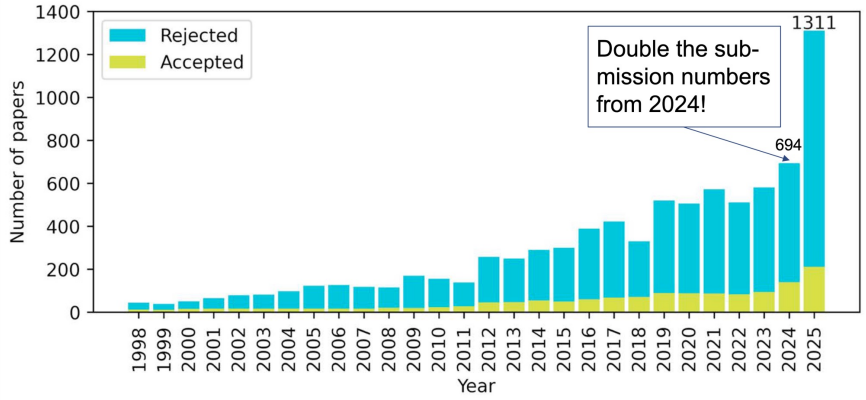
# Challenge Growing Number of Submissions

- **Are peer reviews still scaling?**
  - Growing number of researchers
    - Implies also an increase of submissions
  - Publish or perish culture
  - PC sizes are growing in somewhat comparable pace
    - Can we still fulfill the promise of peer reviews, i.e., timely reviews by experts in the field?
      - Emergency PC members, subreviewing, etc.
      - Reviewing fatigue

Any alternatives to this dilemma?




Paper submissions at USENIX Security




Paper submissions at NDSS



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# LLMs in Automation Tasks



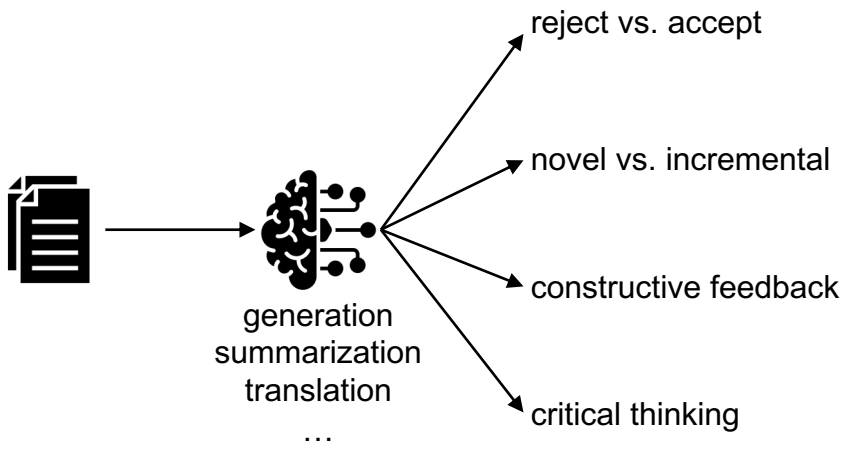
- **LLMs are more often used in various automation tasks**
  - Intrusion and anomaly detection
  - Secure coding support
  - Automated penetration testing
- **Deep natural language understanding**
  - Papers, documentations, …
- **Code synthesis and debugging**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# LLMs in Automation Tasks

- **LLMs are more often used in various automation tasks**
  - Intrusion and anomaly detection
  - Secure coding support
  - Automated penetration testing
- **Deep natural language understanding**
  - Papers, documentations, …
- **Code synthesis and debugging**

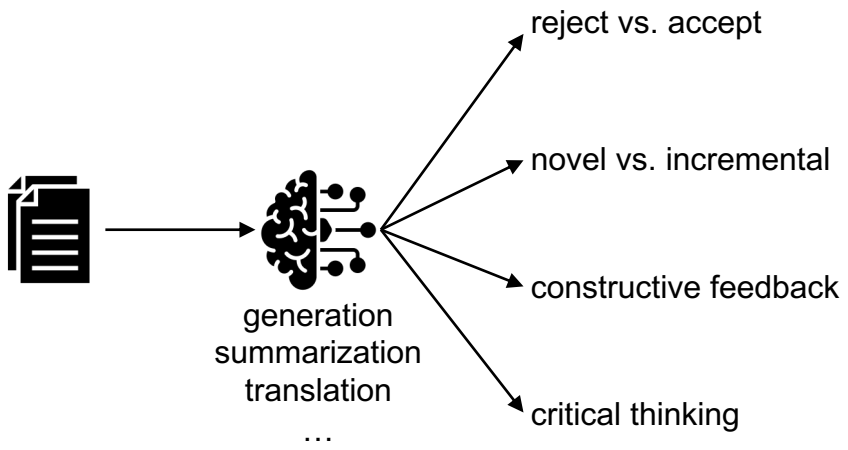





Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# LLMs in Automation Tasks

- **LLMs are more often used in various automation tasks**
  - Intrusion and anomaly detection
  - Secure coding support
  - Automated penetration testing
- **Deep natural language understanding**
  - Papers, documentations, …
- **Code synthesis and debugging**

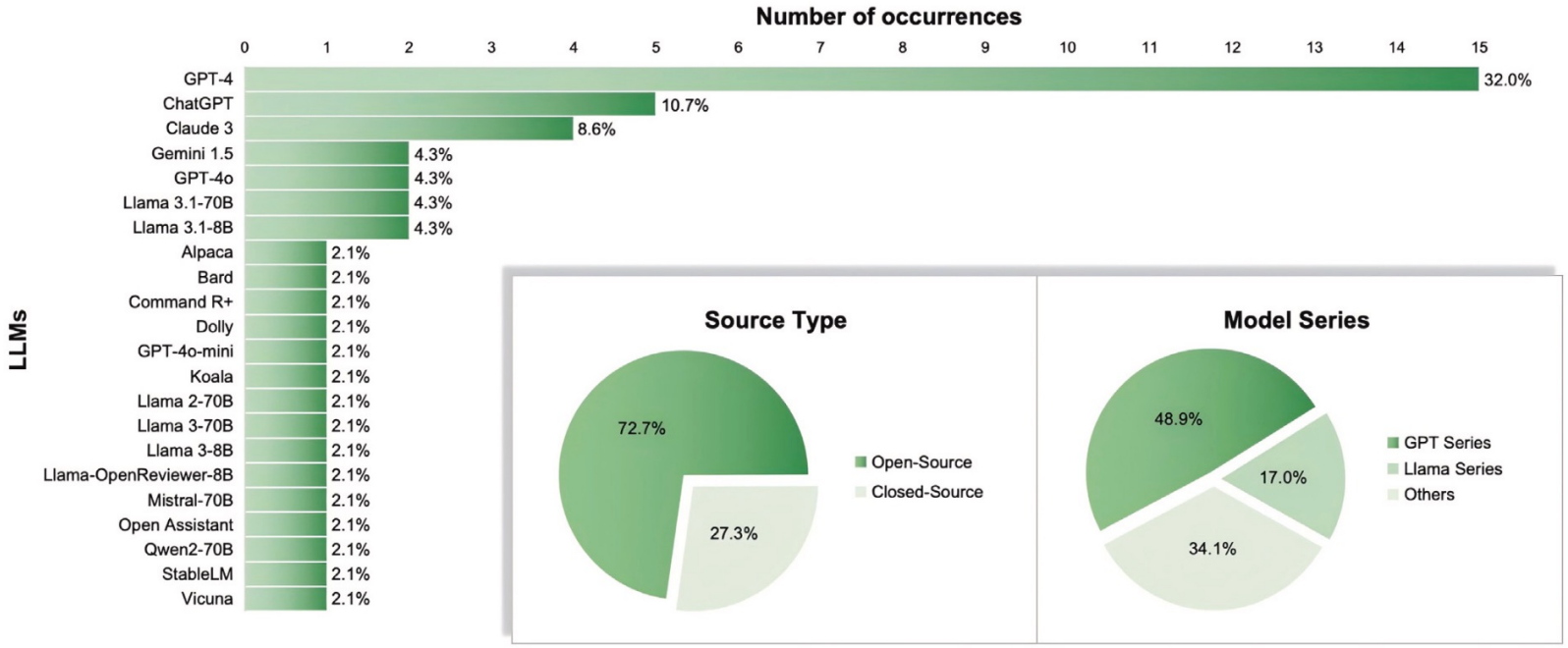



**Can LLMs help in automating peer reviews?**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Related Work on LLMs for Peer Reviews



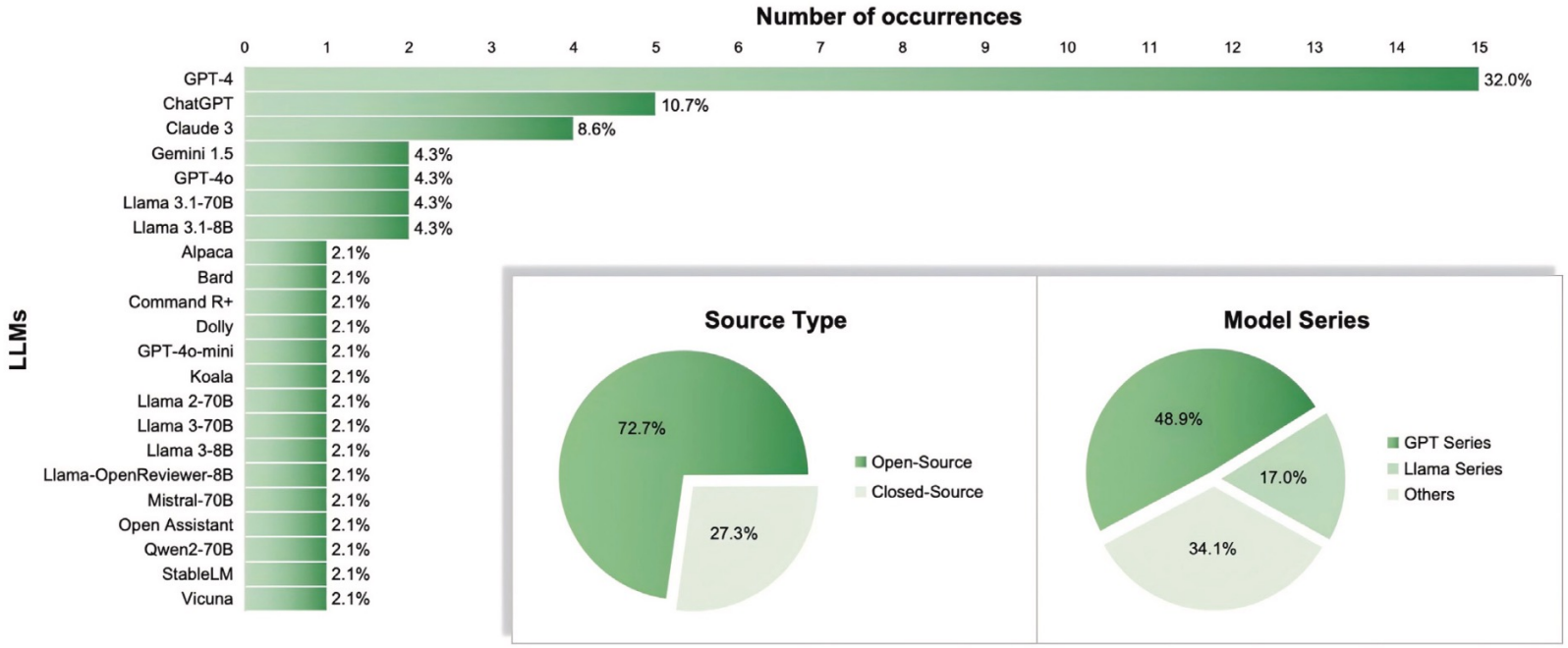

Zhuang et al. (2025) “Large language models for automated scholarly paper review: A survey” Information Fusion, 124

- **Diversity in applied LLMs for peer-reviewing task and focus areas**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Related Work on LLMs for Peer Reviews




Zhuang et al. (2025) “Large language models for automated scholarly paper review: A survey” Information Fusion, 124

- **Diversity in applied LLMs for peer-reviewing task and focus areas**
- **Many open issues remain:** inadequate data, imperfect document parsing, flawed reasoning, …
- **No standardized benchmarks and quality metrics, lack of robustness discussions**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Takeaway on the Suitability of LLMs for Peer Review



- **Significant diversity when applying LLMs for peer review**
  - The wide range of datasets show that many issues remain before practical application
  - Including responsibility/liability questions and how to mitigate poisoning attacks





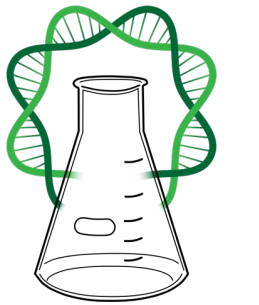
Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Takeaway on the Suitability of LLMs for Peer Review

- **Significant diversity when applying LLMs for peer review**
  - The wide range of datasets show that many issues remain before practical application
  - Including responsibility/liability questions and how to mitigate poisoning attacks



- **Another important angle in recent years is the strive for open science**
  - Official (peer-reviewed) assessments of reproducibility and reusability improve its reputation
  - Peer reviewing open science (research artifacts) differs vastly from “simple” paper reviewing





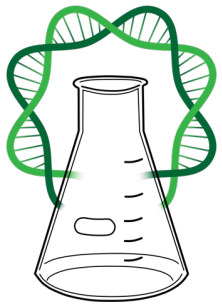

Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Takeaway on the Suitability of LLMs for Peer Review

- **Significant diversity when applying LLMs for peer review**
  - The wide range of datasets show that many issues remain before practical application
  - Including responsibility/liability questions and how to mitigate poisoning attacks



- **Another important angle in recent years is the strive for open science**
  - Official (peer-reviewed) assessments of reproducibility and reusability improve its reputation
  - Peer reviewing open science (research artifacts) differs vastly from “simple” paper reviewing




Which implications does this difference have on the suitability of LLMs for reviewing?



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Excursus: What are YOU contributing these days?

- **The system builds on everyone to chip in**
  - “Golden” rule is a ratio of 3:1 (reviews to papers)



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Excursus: What are YOU contributing these days?

- **The system builds on everyone to chip in**
  - "Golden" rule is a ratio of 3:1 (reviews to papers)
    - Disregards co-authors (positive impact)
    - Disregards rejections and revisions (negative impact)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Excursus: What are YOU contributing these days?

- **The system builds on everyone to chip in**
  - "Golden" rule is a ratio of 3:1 (reviews to papers)
    - Disregards co-authors (positive impact)
    - Disregards rejections and revisions (negative impact)

**Peer Review Metrics**

| 184 | 48 | 2.6:1 |
|---|---|---|
| Verified Peer Reviews | Verified Peer Reviews (Last 12 Months) | Peer Review to Publication Ratio |
| Median: 4<br>99th percentile | Median: 0<br>99th percentile | Median: 0.3:1 |

My Web of Science profile (reviewing statistics)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Excursus: What are YOU contributing these days?

- **The system builds on everyone to chip in**
  - "Golden" rule is a ratio of 3:1 (reviews to papers)
    - Disregards co-authors (positive impact)
    - Disregards rejections and revisions (negative impact)
  - Ideally, for every review you get, you write one yourself
    - Remember: We are looking for constructive criticism

**Peer Review Metrics**

| 184 | 48 | 2.6:1 |
|---|---|---|
| Verified Peer Reviews | Verified Peer Reviews (Last 12 Months) | Peer Review to Publication Ratio |
| Median: 4<br>99th percentile | Median: 0<br>99th percentile | Median: 0.3:1 |

My Web of Science profile (reviewing statistics)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Excursus: What are YOU contributing these days?

- **The system builds on everyone to chip in**
  - "Golden" rule is a ratio of 3:1 (reviews to papers)
    - Disregards co-authors (positive impact)
    - Disregards rejections and revisions (negative impact)
  - Ideally, for every review you get, you write one yourself
    - Remember: We are looking for constructive criticism




My Web of Science profile (reviewing statistics)

1) **Who has submitted research papers for peer review so far?**
2) **Who contributed to the community as a reviewer in 2025?**
3) **Who has already served on a conference PC?**
4) **Who has participated in an artifact evaluation PC?**


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Artifact Evaluation at Cybersecurity Conferences



- **Reproducibility bridges research and production**
  - Validate results
  - Enable reuse and comparison



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Artifact Evaluation at Cybersecurity Conferences



- **Reproducibility bridges research and production**
  - Validate results
  - Enable reuse and comparison
- **More conferences introduce AE**
  - Award badges (*available, functional, reproduced*)
  - Sometimes even mandatory (e.g., USENIX Sec, JSys, …)



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Artifact Evaluation at Cybersecurity Conferences



- **Reproducibility bridges research and production**
  - Validate results
  - Enable reuse and comparison
- **More conferences introduce AE**
  - Award badges (*available, functional, reproduced*)
  - Sometimes even mandatory (e.g., USENIX Sec, JSys, …)
- **Extremely high human effort**
  - ACM CCS 25: **131** artifacts, **125** AEC members, **3.4** reviews
  - NDSS 26: **116** artifacts, **128** AEC members, **2** or **3** reviews



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Artifact Evaluation at Cybersecurity Conferences



- **Reproducibility bridges research and production**
  - Validate results
  - Enable reuse and comparison
- **More conferences introduce AE**
  - Award badges (*available, functional, reproduced*)
  - Sometimes even mandatory (e.g., USENIX Sec, JSys, …)
- **Extremely high human effort**
  - ACM CCS 25: **131** artifacts, **125** AEC members, **3.4** reviews
  - NDSS 26: **116** artifacts, **128** AEC members, **2** or **3** reviews
- **Increasing complexity of artifacts and their setup requirements**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Artifact Evaluation at Cybersecurity Conferences

- **Reproducibility bridges research and production**
  - Validate results
  - Enable reuse and comparison
- **More conferences introduce AE**
  - Award badges (*available, functional, reproduced*)
  - Sometimes even mandatory (e.g., USENIX Sec, JSys, …)
- **Extremely high human effort**
  - ACM CCS 25: **131** artifacts, **125** AEC members, **3.4** reviews
  - NDSS 26: **116** artifacts, **128** AEC members, **2** or **3** reviews
- **Increasing complexity of artifacts and their setup requirements**




**The little brother of paper peer reviews…**

- Frequent calls for self-nomination
- Oftentimes targeted at PhD students

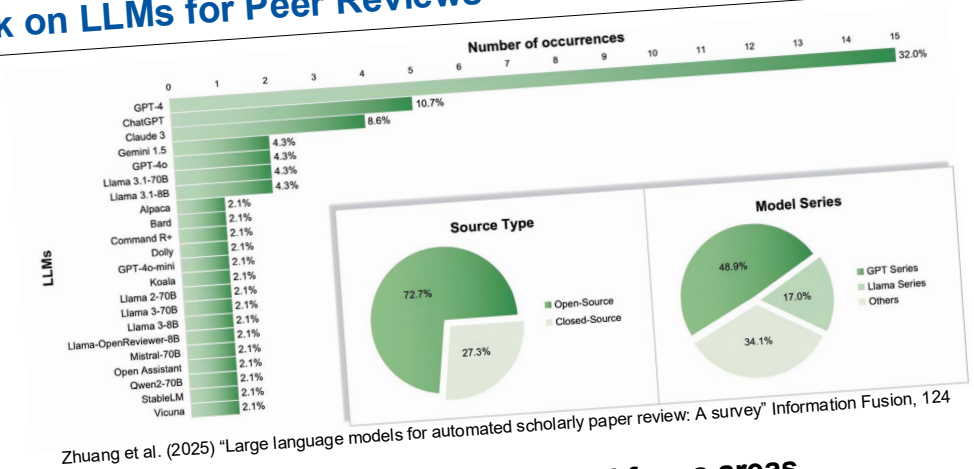
**AE—just the unpleasant side task?**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# LLMs in Peer-Reviewing Tasks

## Related Work on LLMs for Peer Reviews




Zhuang et al. (2025) “Large language models for automated scholarly paper review: A survey” Information Fusion, 124

- **Diversity in applied LLMs for peer-reviewing task and focus areas**
- **Many open issues remain:** inadequate data, imperfect document parsing, flawed reasoning, …
- **No standardized benchmarks and quality metrics, lack of robustness discussions**

Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

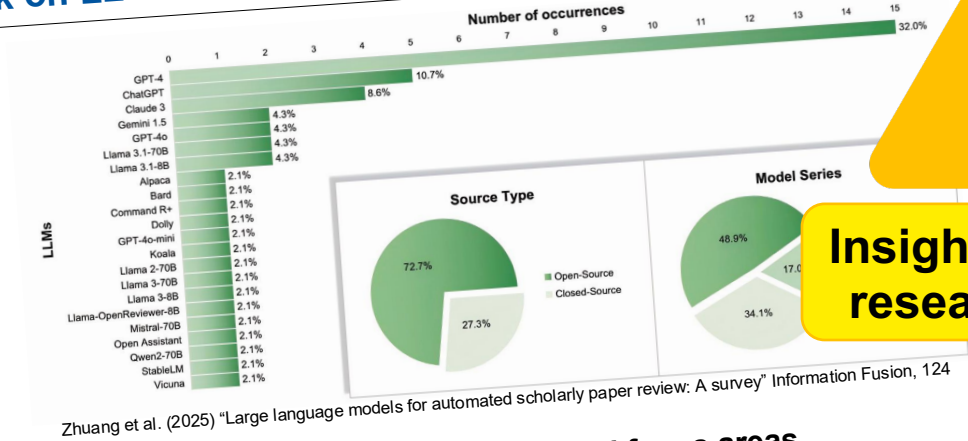




Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# LLMs in Peer-Reviewing Tasks

## Related Work on LLMs for Peer Reviews



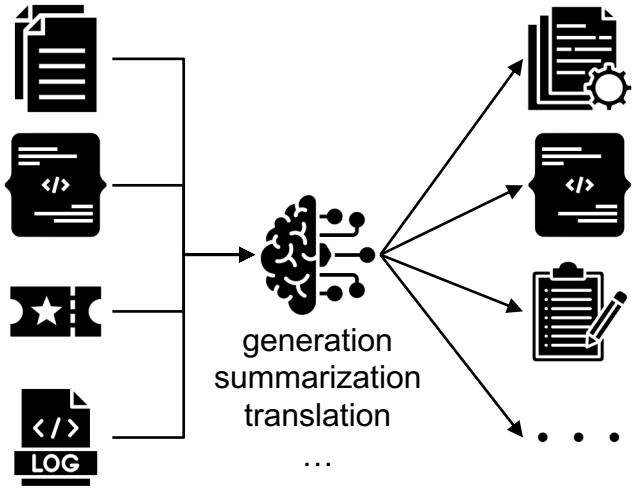

Zhuang et al. (2025) "Large language models for automated scholarly paper review: A survey" Information Fusion, 124

**Insights limited to research papers**

- **Diversity in applied LLMs for peer-reviewing task and focus areas**
- **Many open issues remain:** inadequate data, imperfect document parsing, flawed reasoning, …
- **No standardized benchmarks and quality metrics, lack of robustness discussions**

Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

7

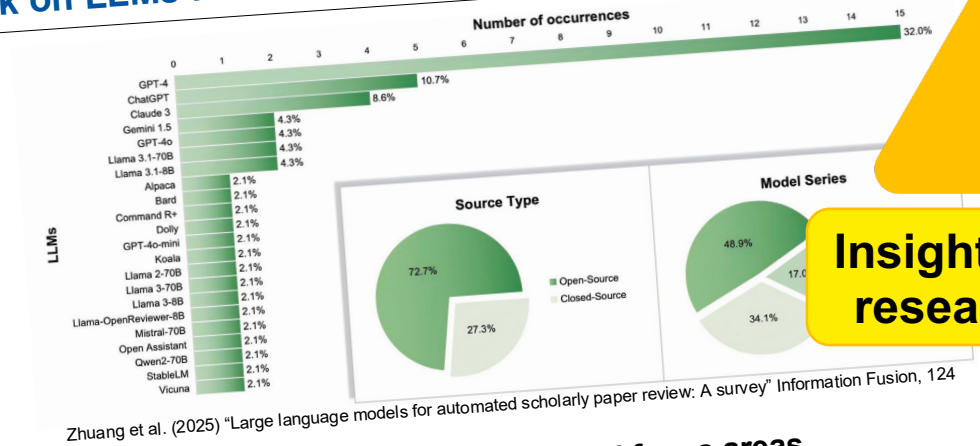





Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# LLMs in Peer-Reviewing Tasks

## Related Work on LLMs for Peer Reviews



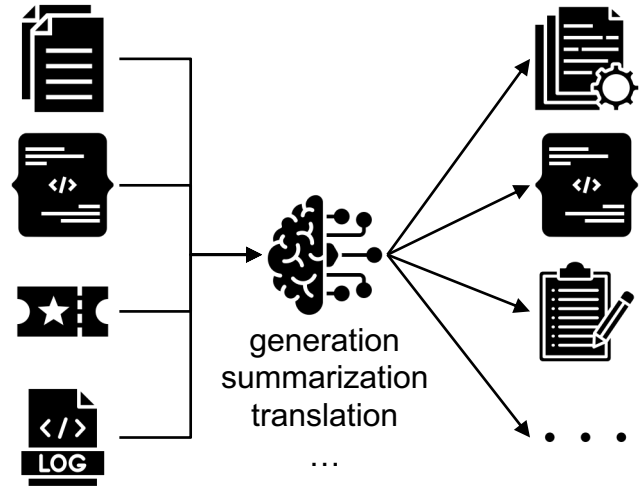

Zhuang et al. (2025) "Large language models for automated scholarly paper review: A survey" Information Fusion, 124

**Insights limited to research papers**

- **Diversity in applied LLMs for peer-reviewing task and focus areas**
- **Many open issues remain:** inadequate data, imperfect document parsing, flawed reasoning, …
- **No standardized benchmarks and quality metrics, lack of robustness discussions**

Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

7




**How can LLMs help automate key artifact evaluation tasks?**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Olszewski and Arp Studies

- **Olszewski et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Olszewski and Arp Studies

- **Olszewski et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results



**Only a minority of artifacts are actually reproducible**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Olszewski and Arp Studies

- **Olszewski et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results

- **Arp et al., 2022:**
  **"Dos and Don'ts of Machine Learning in Computer Security"**
  - Review **30** papers to identify experimental and design pitfalls
  - Sampling bias affects **90%** of the papers
  - **65%** use an inappropriate threat model
  - Very relevant for peer review

**Only a minority of artifacts are actually reproducible**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Olszewski and Arp Studies

- **Olszewski et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results

- **Arp et al., 2022:**
  **"Dos and Don'ts of Machine Learning in Computer Security"**
  - Review **30** papers to identify experimental and design pitfalls
  - Sampling bias affects **90%** of the papers
  - **65%** use an inappropriate threat model
  - Very relevant for peer review

**Only a minority of artifacts are actually reproducible**

**Many papers are affected by pitfalls overestimating their performance**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Olszewski and Arp Studies

I attended the presentation

- **Olszewski et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results

- **Arp et al., 2022:**
  **"Dos and Don'ts of Machine Learning in Computer Security"**
  - Review **30** papers to identify experimental and design pitfalls
  - Sampling bias affects **90%** of the papers
  - **65%** use an inappropriate threat model
  - Very relevant for peer review

**Only a minority of artifacts are actually reproducible**

**Many papers are affected by pitfalls overestimating their performance**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

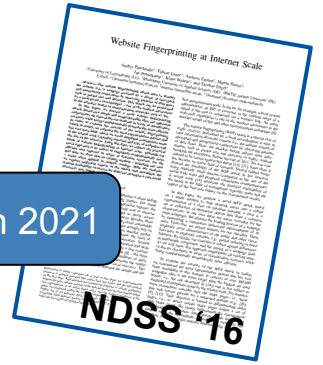
# OLSZEWSKI and ARP Studies

I attended the presentation

- **OLSZEWSKI et al., 2023:**
  **"Get in Researchers; We're Measuring Reproducibility"**
  - Manually assess reproducibility levels of cybersecurity artifact submissions
    - Consider top-tier conferences, 2013-2022
  - **744** papers, only **298** include artifacts
  - **57%** of artifacts include a readme file
  - **46%** of artifacts are runnable
  - **20%** of artifacts reproduce the paper's results

- **ARP et al., 2022:**
  **"Dos and Don'ts of Machine Learning in Computer Security"**
  - Review **30** papers to identify experimental and design pitfalls
  - Sampling bias affects **90%** of the papers
  - **65%** use an inappropriate threat model
  - Very relevant for peer review

I was a study/survey participant in 2021



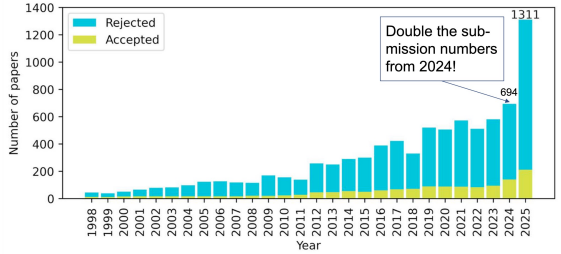

**Only a minority of artifacts are actually reproducible**

**Many papers are affected by pitfalls overestimating their performance**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# An Overview of Open Research Issues



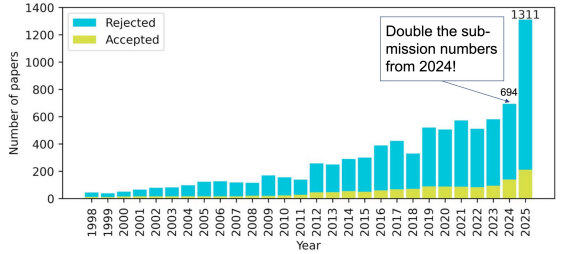

**submission numbers + manual effort**

USENIX Security '26 Preliminary Call for Artifacts



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# An Overview of Open Research Issues




**submission numbers + manual effort**

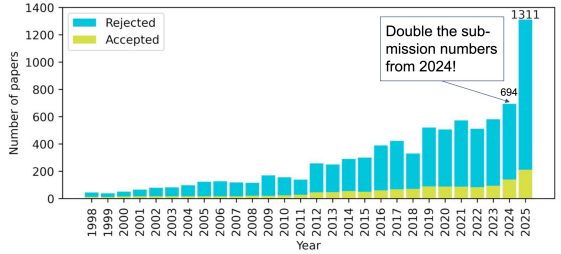
USENIX Security '26 Preliminary Call for Artifacts

**Issue 1:** *Conventional AE processes no longer scale with rising submission rates and the diversity in utilized software and hardware stacks*



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

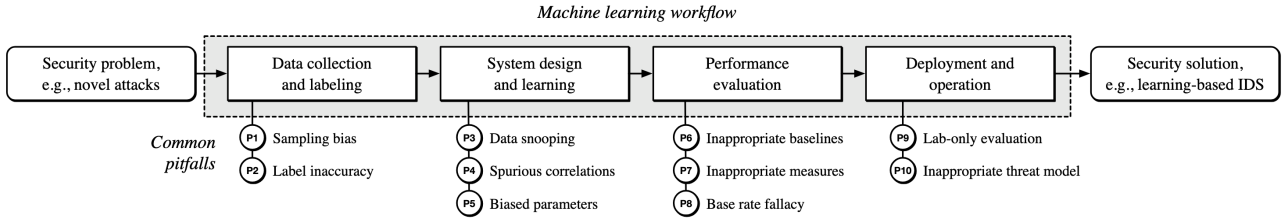
# An Overview of Open Research Issues




**submission numbers + manual effort**

USENIX Security '26 Preliminary Call for Artifacts

**Issue 1:** *Conventional AE processes no longer scale with rising submission rates and the diversity in utilized software and hardware stacks*

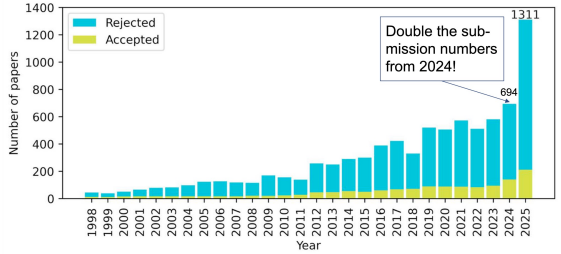



**reproducibility checks seldomly validate the methodology**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

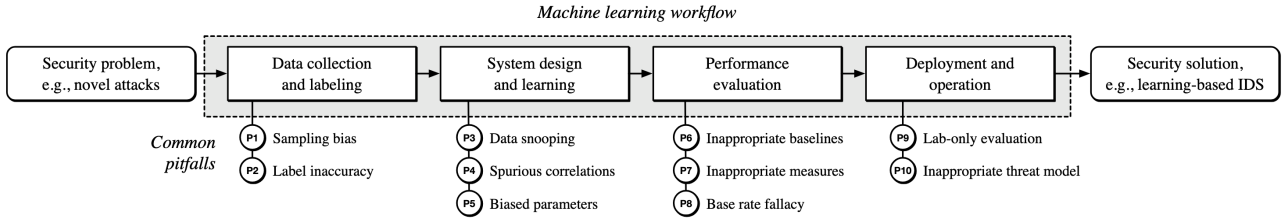
# An Overview of Open Research Issues




**submission numbers + manual effort**

USENIX Security '26 Preliminary Call for Artifacts

**Issue 1:** *Conventional AE processes no longer scale with rising submission rates and the diversity in utilized software and hardware stacks*



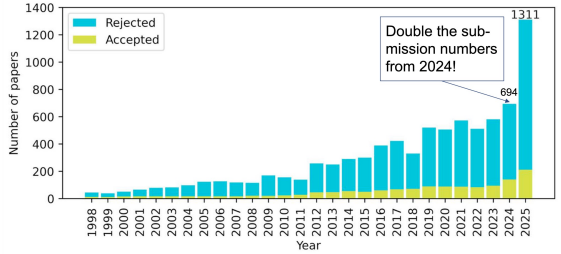

**reproducibility checks seldomly validate the methodology**

**Issue 2:** *Detecting methodological flaws in a study is vital to determine its true contribution; however, these flaws are often hard to detect as part of standard reproducibility checks*



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

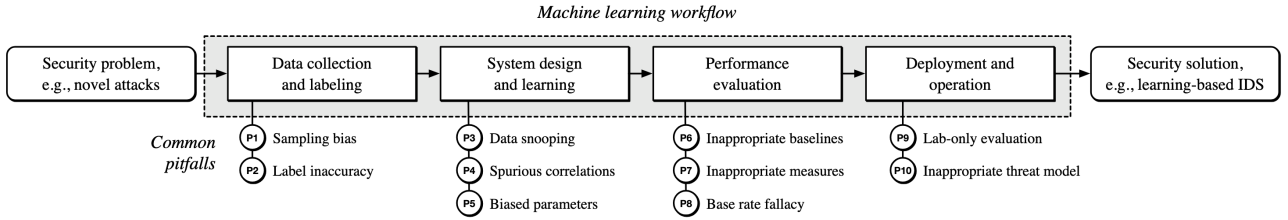
# An Overview of Open Research Issues




**submission numbers + manual effort**

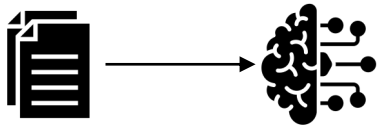
USENIX Security '26 Preliminary Call for Artifacts

**Issue 1:** *Conventional AE processes no longer scale with rising submission rates and the diversity in utilized software and hardware stacks*




**reproducibility checks seldomly validate the methodology**

**Issue 2:** *Detecting methodological flaws in a study is vital to determine its true contribution; however, these flaws are often hard to detect as part of standard reproducibility checks*



**explored but with room for potential**



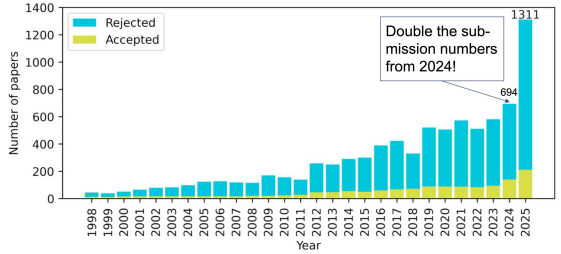

**benefits unclear**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

COMSYS | RWTH AACHEN UNIVERSITY

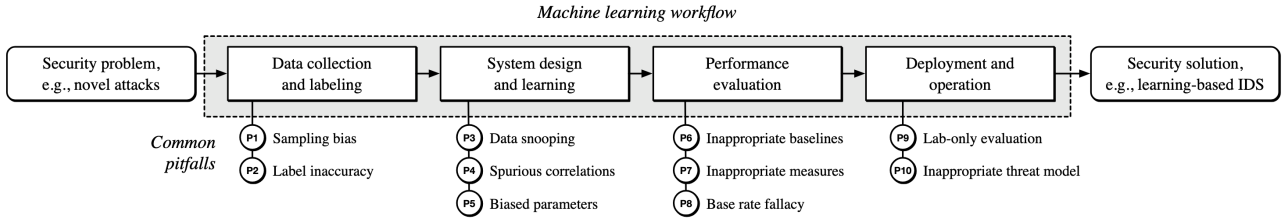
# An Overview of Open Research Issues




**submission numbers + manual effort**

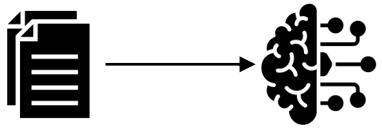
USENIX Security '26 Preliminary Call for Artifacts

**Issue 1:** *Conventional AE processes no longer scale with rising submission rates and the diversity in utilized software and hardware stacks*




**reproducibility checks seldomly validate the methodology**

**Issue 2:** *Detecting methodological flaws in a study is vital to determine its true contribution; however, these flaws are often hard to detect as part of standard reproducibility checks*



**explored but with room for potential**




**benefits unclear**

**Issue 3:** *The utility of AI for assessing the reproducibility of proposed concepts (e.g., as part of AE) remains underexplored*



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

COMSYS | RWTH AACHEN UNIVERSITY

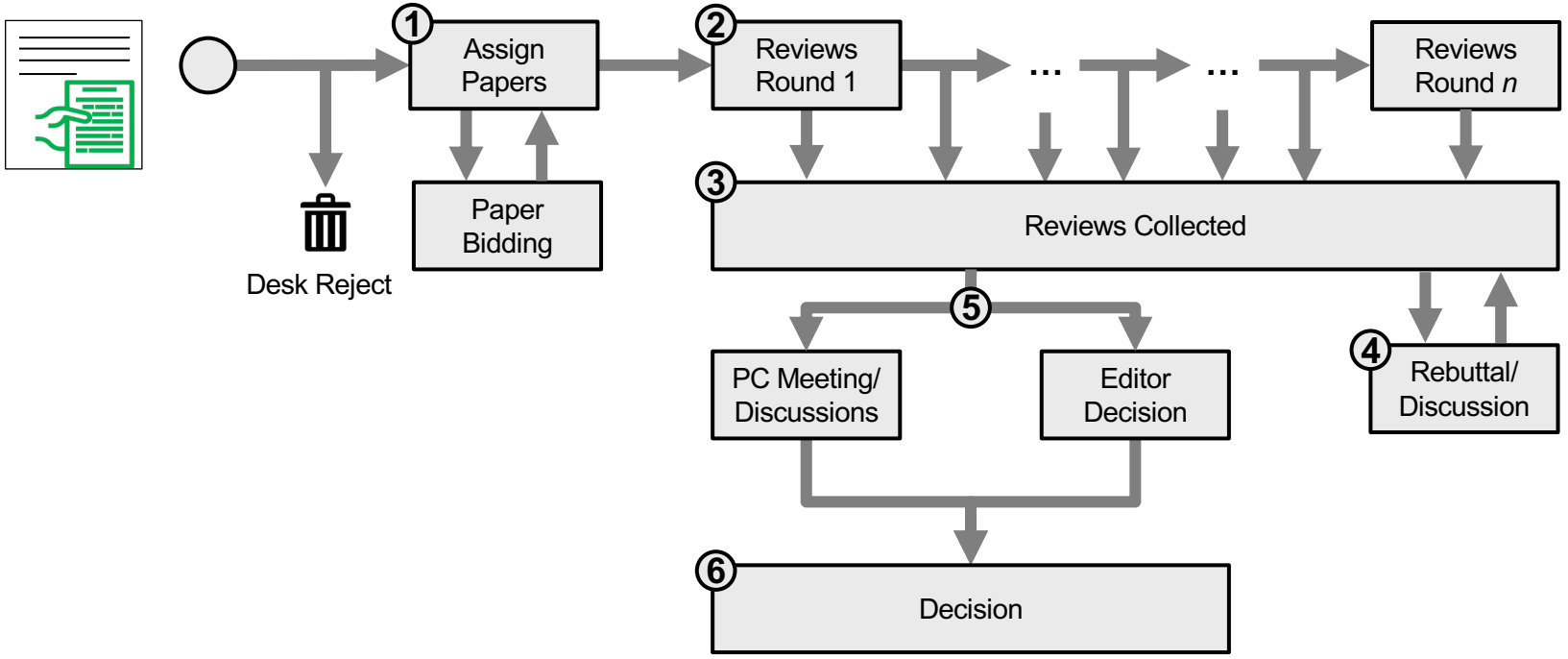
# Interplay within the Research Landscape





Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

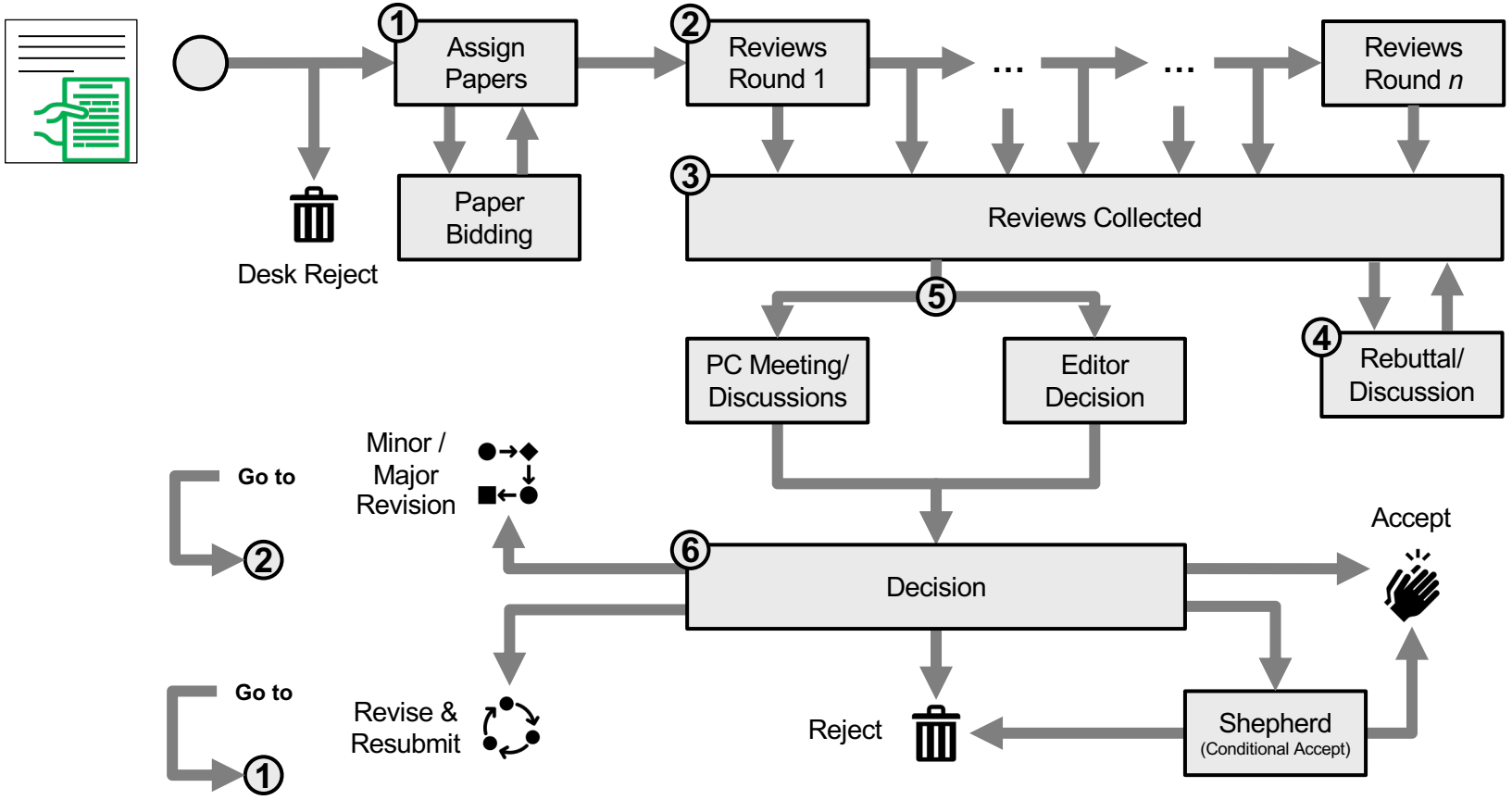
# Interplay within the Research Landscape






Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

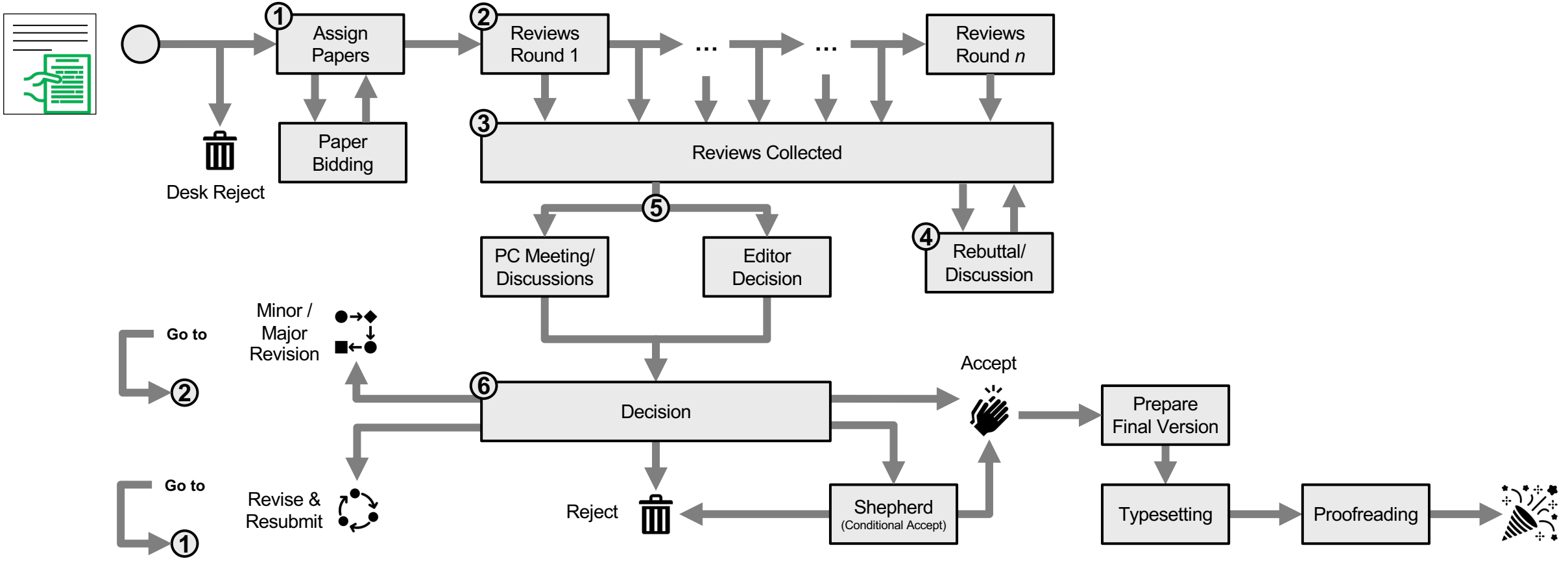
# Interplay within the Research Landscape






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

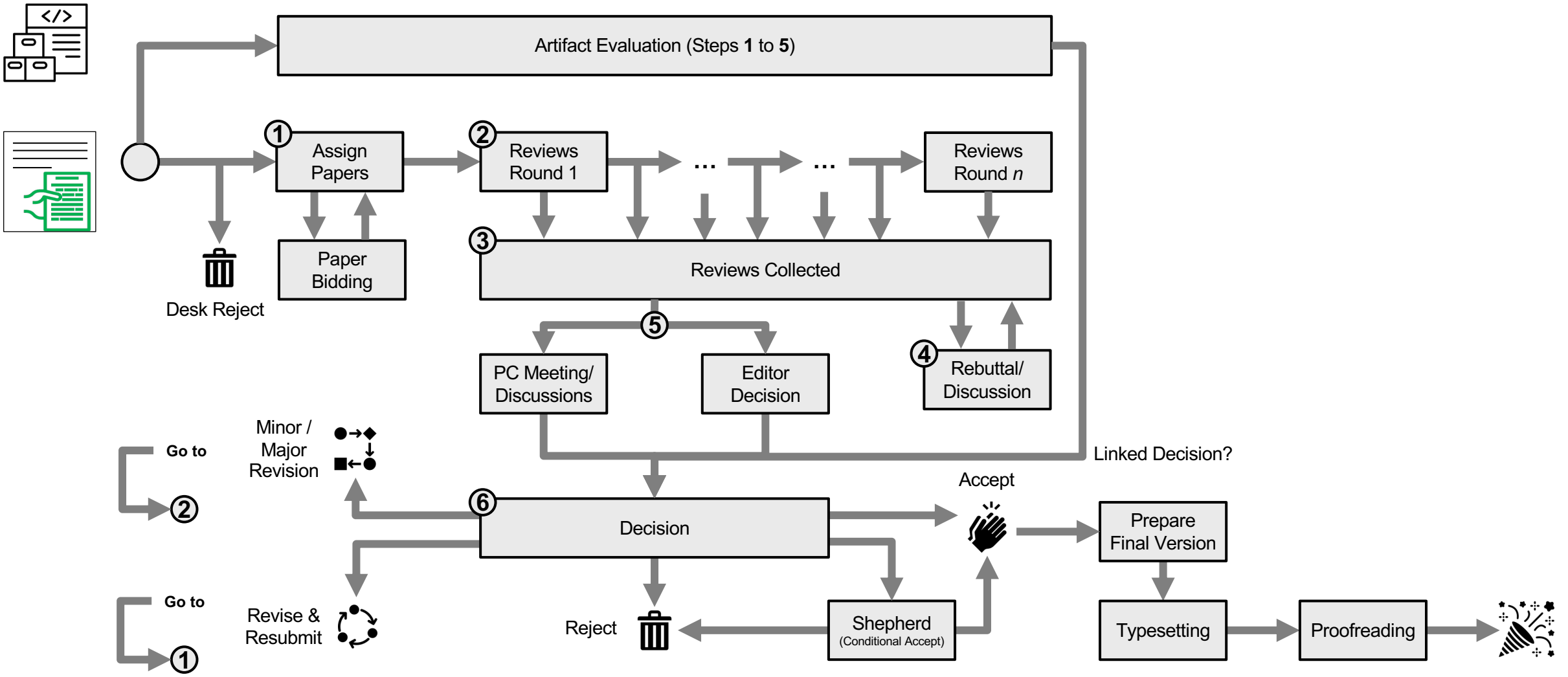
# Interplay within the Research Landscape






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

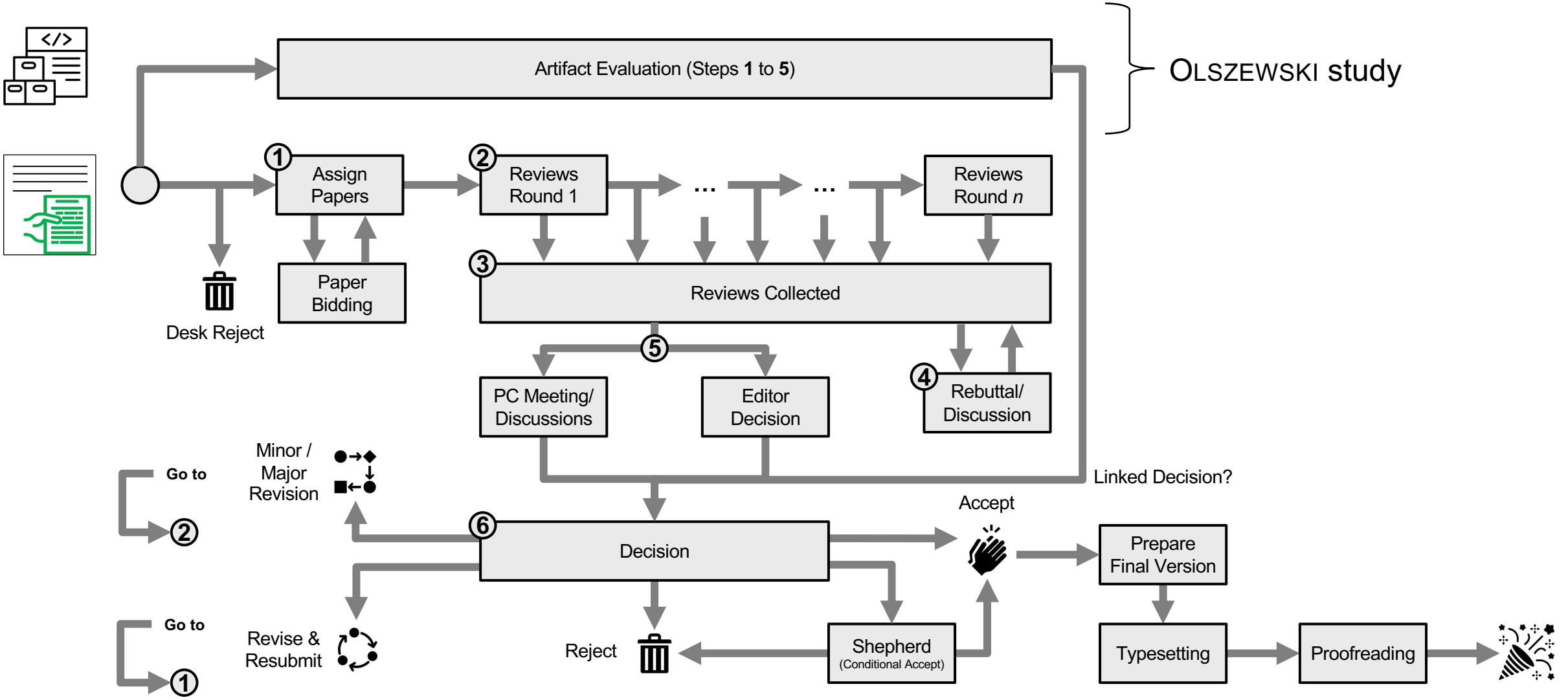
# Interplay within the Research Landscape






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

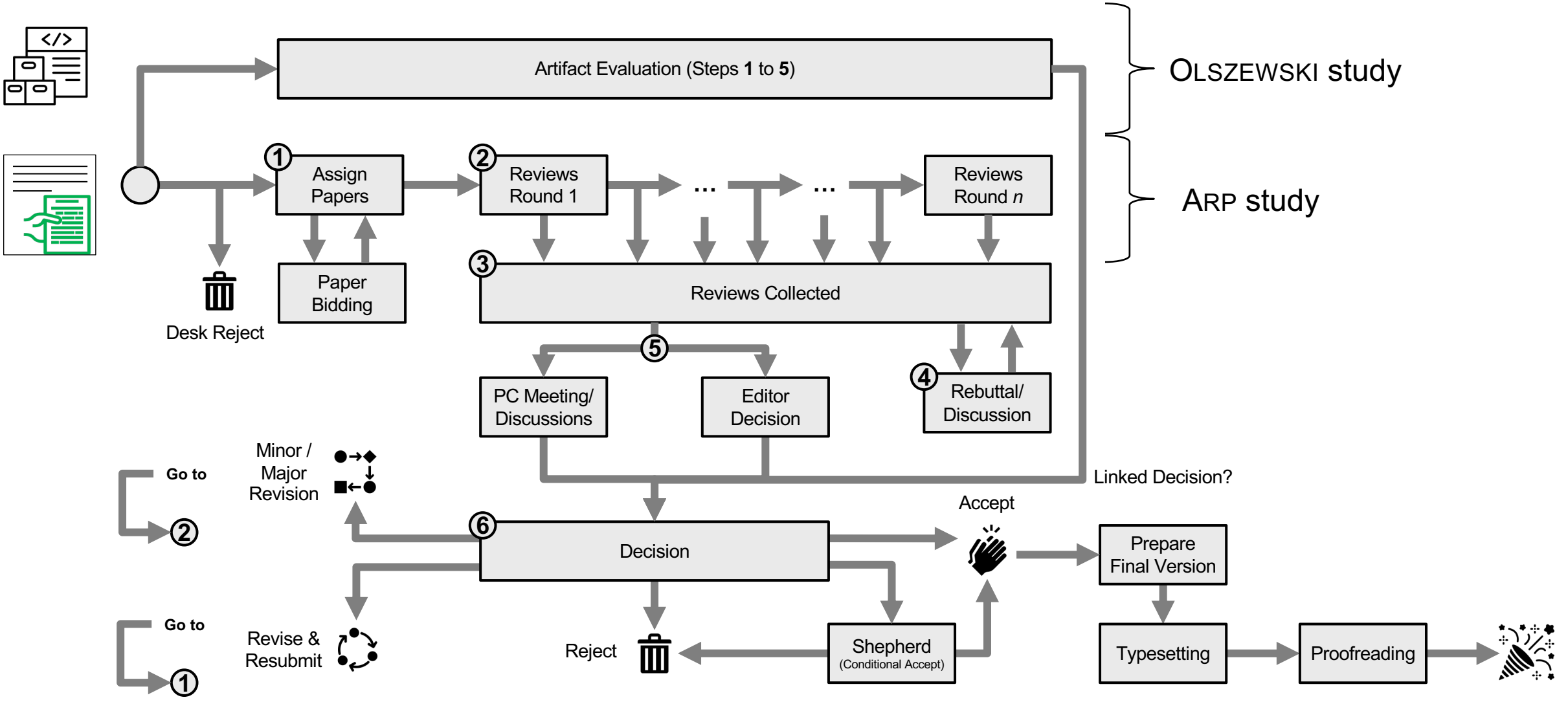
# Interplay within the Research Landscape






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

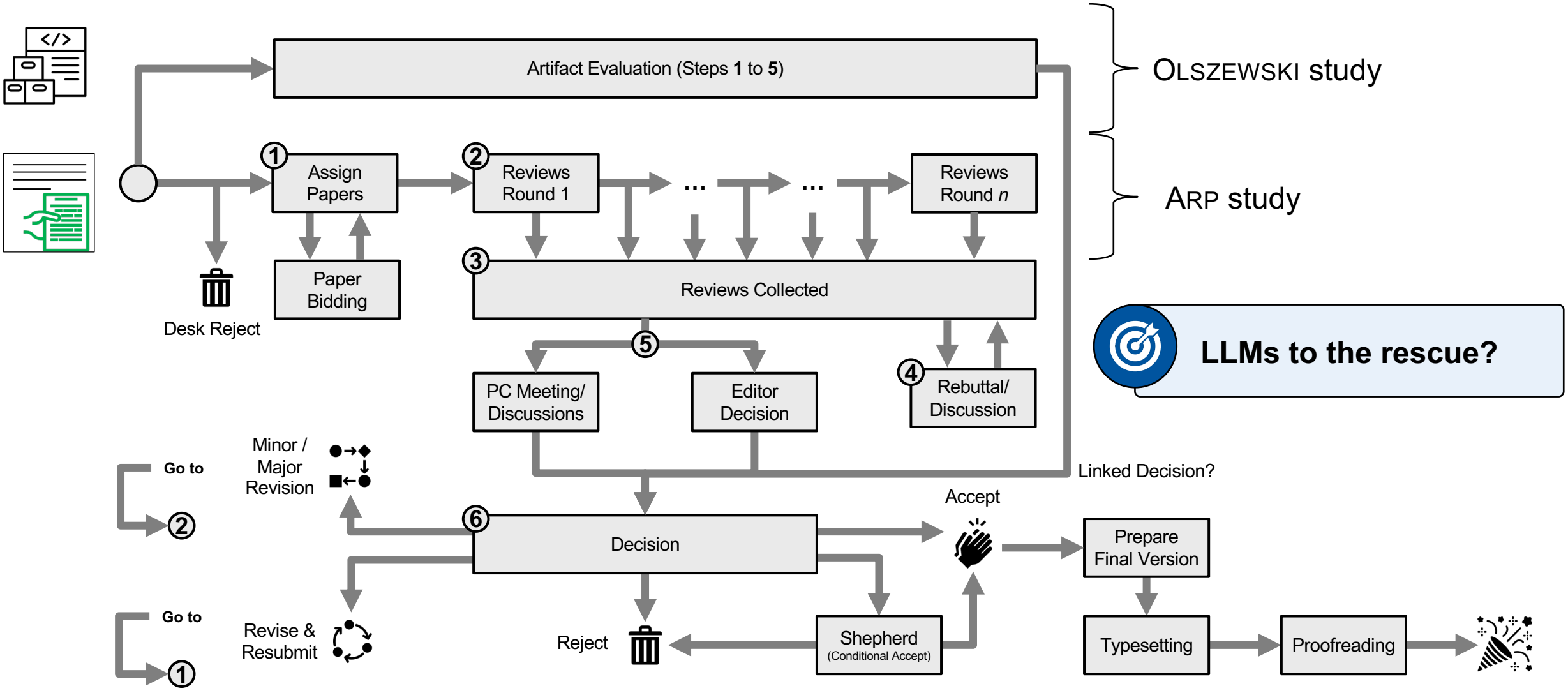
# Interplay within the Research Landscape






Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Interplay within the Research Landscape



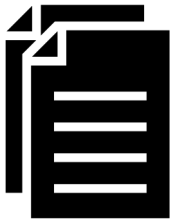



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Three-Step Toolkit

RATE

- **Analyze paper text**
- **Compute reproducibility score**



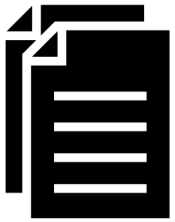


Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Three-Step Toolkit

## RATE

- **Analyze paper text**
- **Compute reproducibility score**



## PREPARE

- **Analyze code repository**
- **Follow Readme instructions**



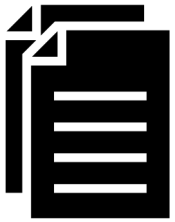


Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Three-Step Toolkit

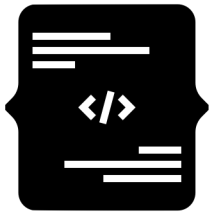
RATE

- **Analyze paper text**
- **Compute reproducibility score**



PREPARE

- **Analyze code repository**
- **Follow Readme instructions**



OLSZEWSKI study

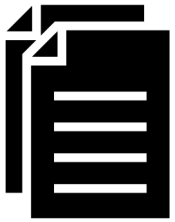


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Three-Step Toolkit

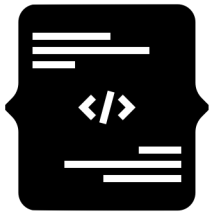
## Rate

- **Analyze paper text**
- **Compute reproducibility score**



## Prepare

- **Analyze code repository**
- **Follow Readme instructions**



Olszewski study

## Assess

- **Analyze paper text**
- **Detect pitfalls**





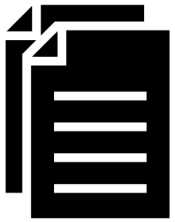
Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Three-Step Toolkit



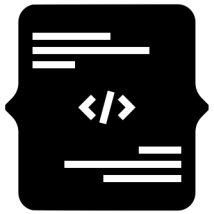
RATE

- **Analyze paper text**
- **Compute reproducibility score**



PREPARE

- **Analyze code repository**
- **Follow Readme instructions**



OLSZEWSKI study

ASSESS

- **Analyze paper text**
- **Detect pitfalls**



ARP study



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]

Filter non-reproducible papers early



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]

reproducible



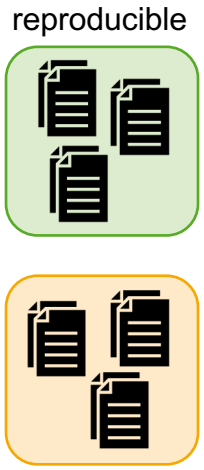
Filter non-reproducible papers early



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



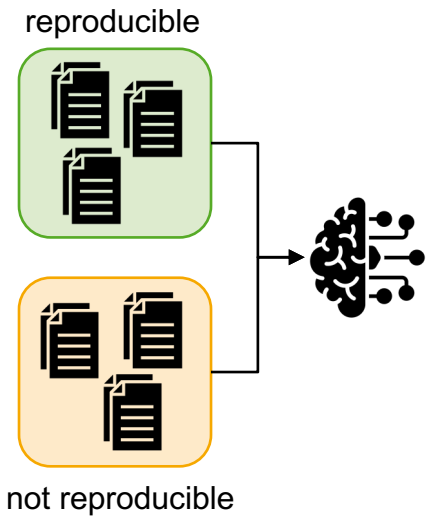

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



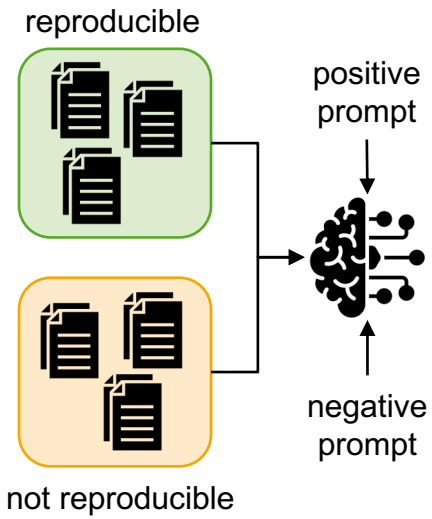

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



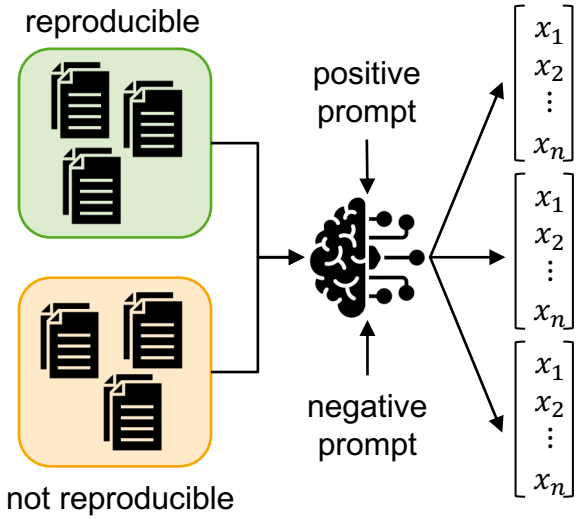

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



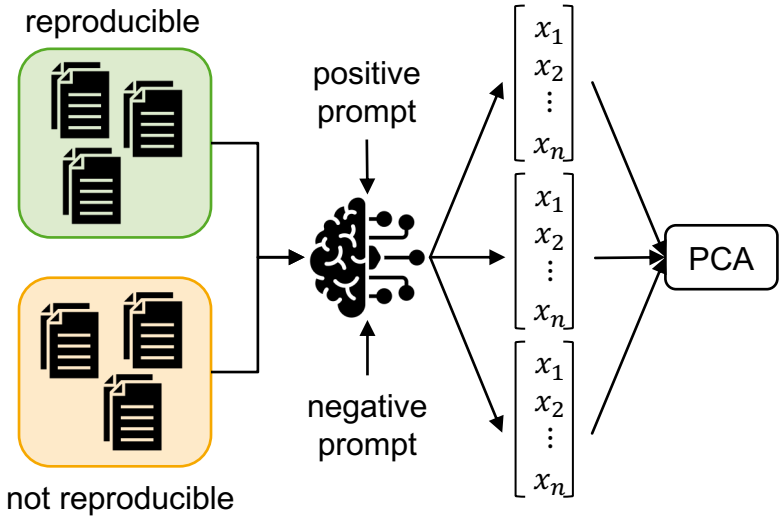

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



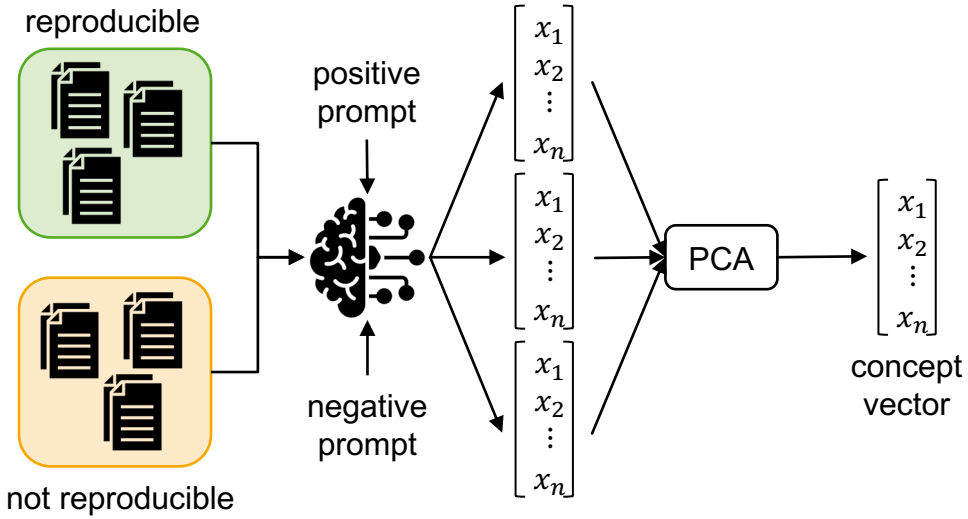

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



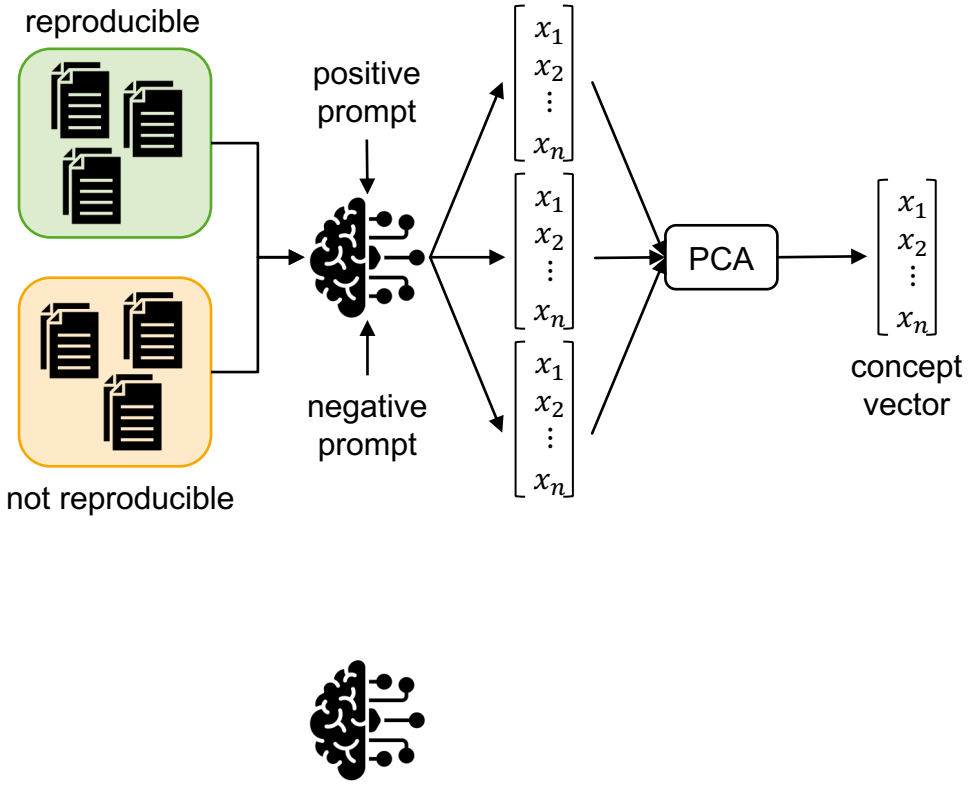

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



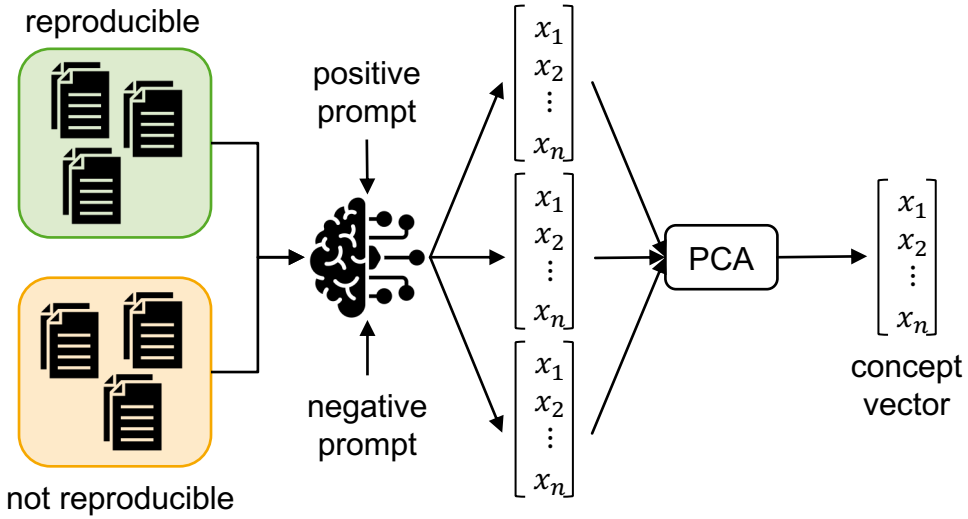

Filter non-reproducible papers early

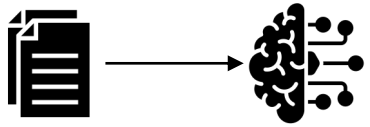


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



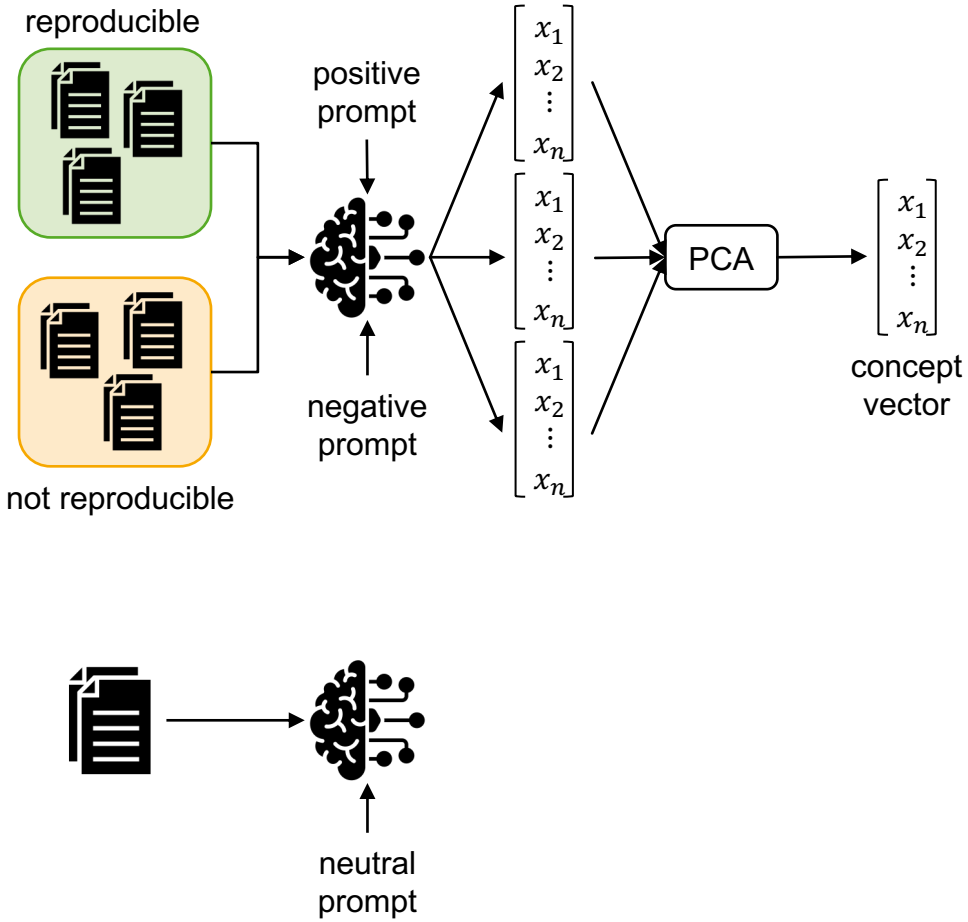
Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



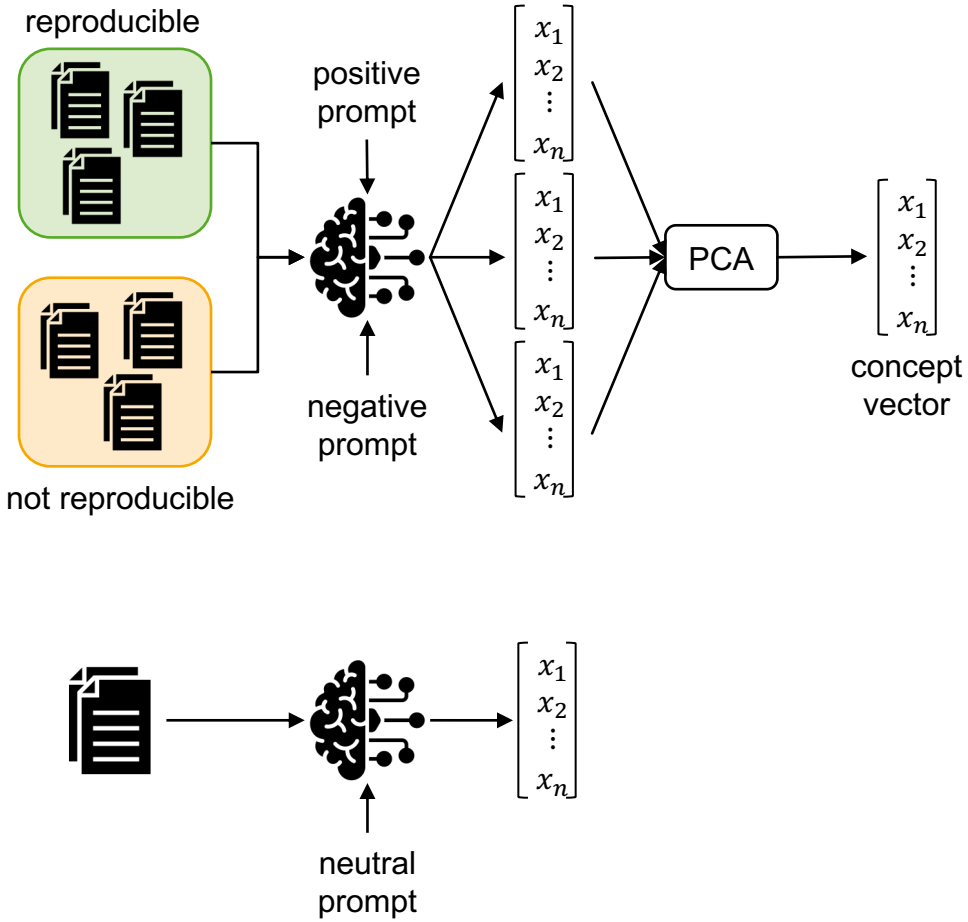

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



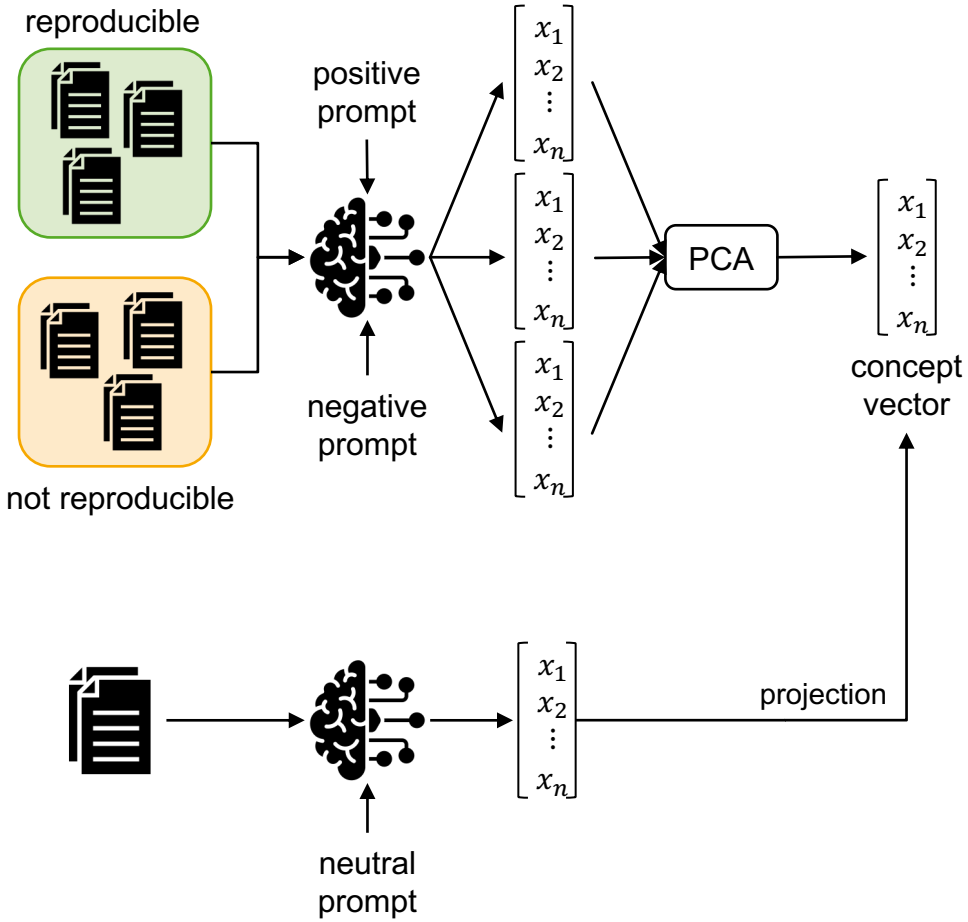

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]



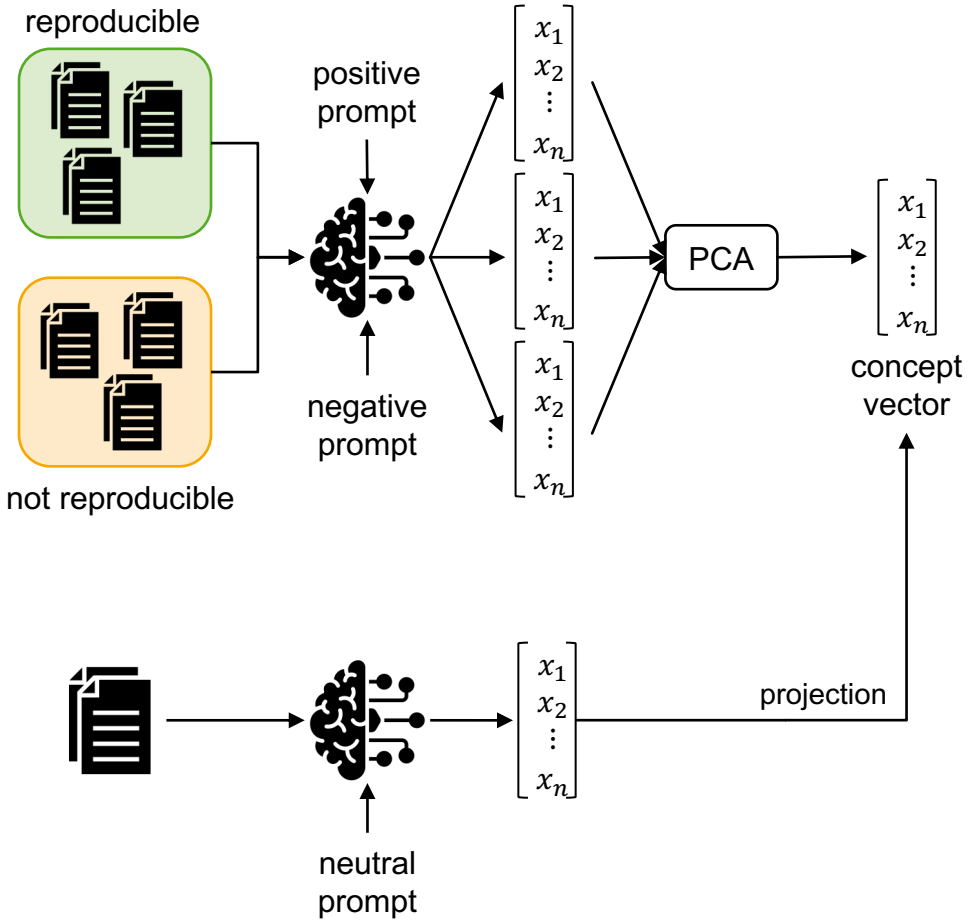

Filter non-reproducible papers early



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]




Filter non-reproducible papers early

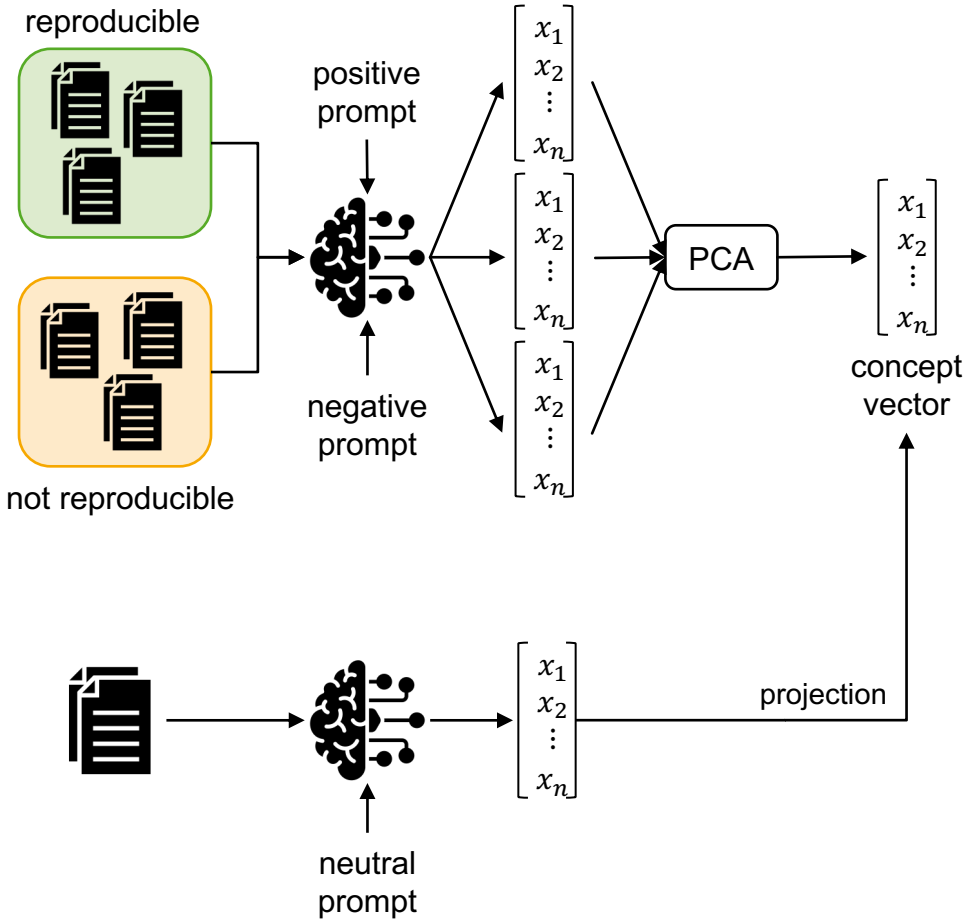
Prioritize recall



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# RATE – Content-Based Reproducibility Score

- **Use LLM internal states** [Yang et al., 2024]




Filter non-reproducible papers early

Prioritize recall

Independent of the model, only require access to final layer

95% true positive rate
5% false negative rate
95% recall

Llama-3.2-3B-Instruct



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code



- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# PREPARE – Autonomously Setting up Code



- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



Set up code execution environments



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



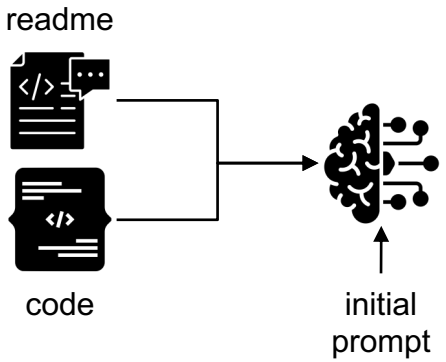

Set up code execution environments



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



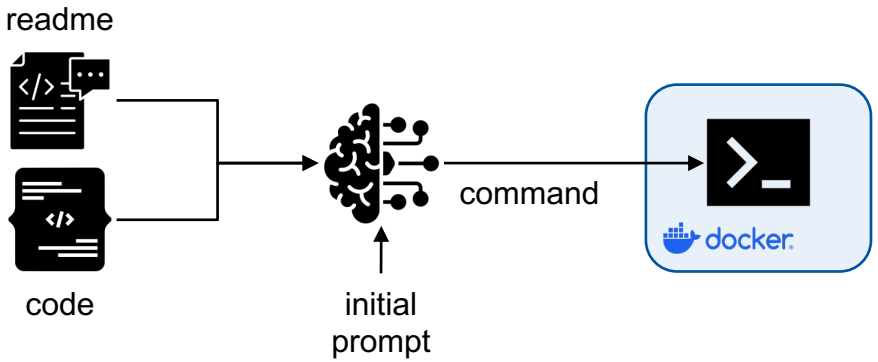

Set up code execution environments



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



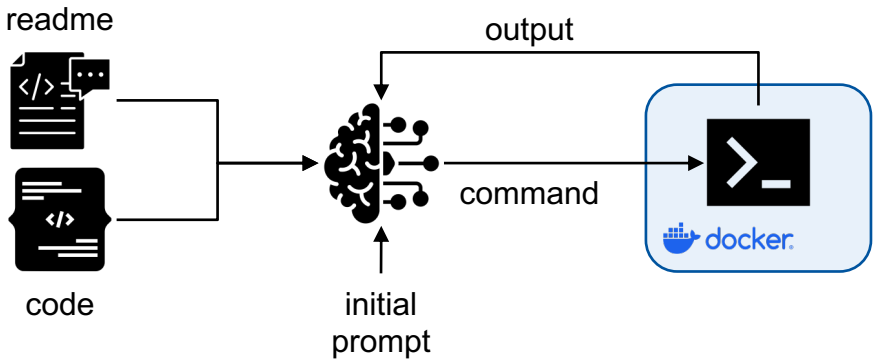

Set up code execution environments



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



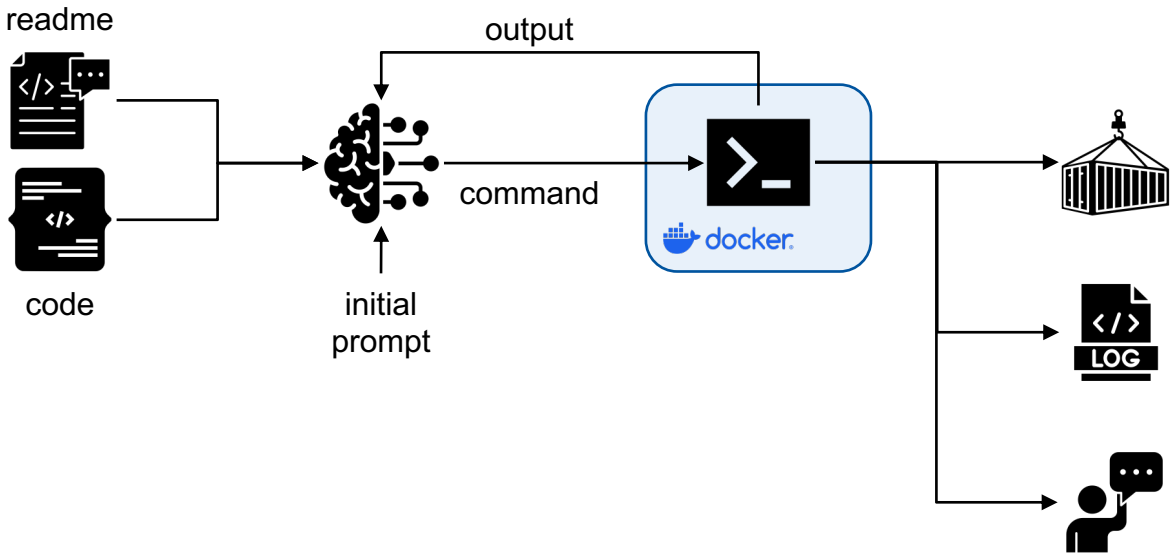

Set up code execution environments



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components



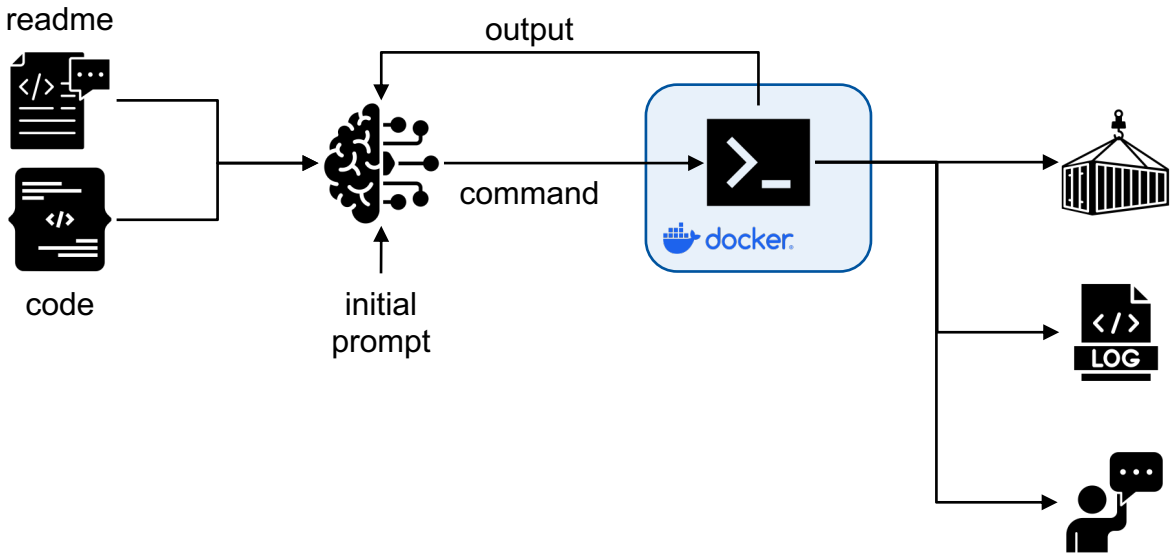

Set up code execution environments



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components




Set up code execution environments

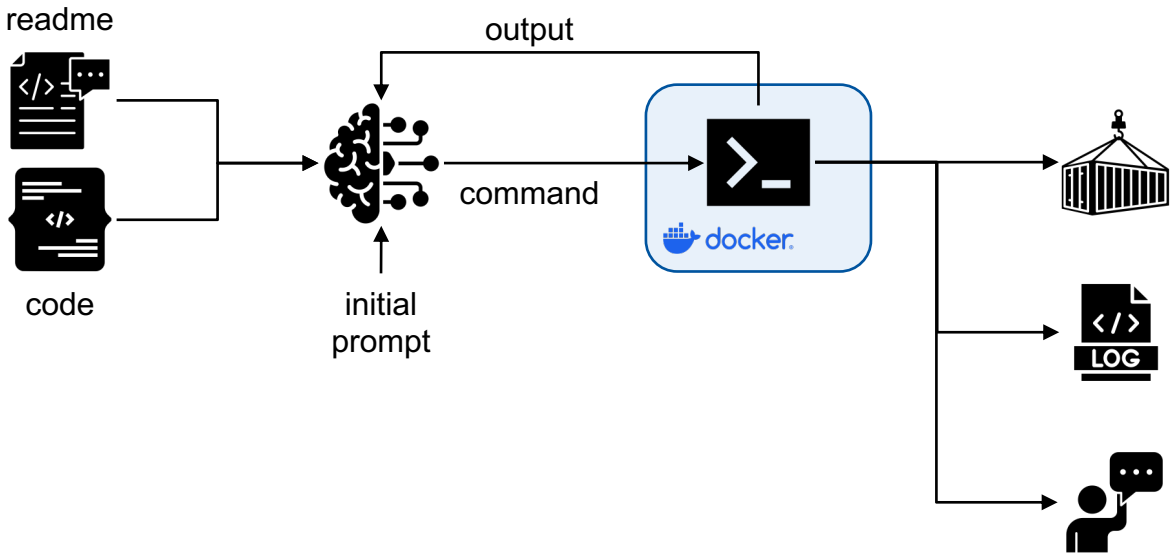
Prioritize accuracy



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# PREPARE – Autonomously Setting up Code

- **LLM agent generating terminal commands**
  - Parse readme file
  - Clone repository, install dependencies
  - Compile and run code components




Set up code execution environments

Prioritize accuracy

Full isolation per artifact via container environments

28% true positive rate
80% true negative rate
66% accuracy

GPT-4o-mini



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# ASSESS – Identifying Pitfalls in Contributions



- **Consider 8 pitfalls from ARP et al.**
  - Sampling bias
  - Label inaccuracy
  - Data snooping
  - Spurious correlations
  - Biased parameters
  - Inappropriate baselines
  - Inappropriate measures
  - Lab-only evaluation
  - Inappropriate threat model
  - Base rate fallacy
- **Same technical approach as for RATE**
  - Find concept vector per pitfall
  - Project hidden state onto concept vector



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# ASSESS – Identifying Pitfalls in Contributions



- **Consider 8 pitfalls from ARP et al.**
  - Sampling bias
  - Label inaccuracy
  - Data snooping
  - Spurious correlations
  - Biased parameters
  - Inappropriate baselines
  - Inappropriate measures
  - Lab-only evaluation
  - Inappropriate threat model
  - Base rate fallacy
- **Same technical approach as for RATE**
  - Find concept vector per pitfall
  - Project hidden state onto concept vector



Detect flaws to later watch out for



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# ASSESS – Identifying Pitfalls in Contributions



- **Consider 8 pitfalls from ARP et al.**
  - Sampling bias
  - Label inaccuracy
  - Data snooping
  - Spurious correlations
  - Biased parameters
  - Inappropriate baselines
  - Inappropriate measures
  - Lab-only evaluation
  - Inappropriate threat model
  - Base rate fallacy
- **Same technical approach as for RATE**
  - Find concept vector per pitfall
  - Project hidden state onto concept vector



Detect flaws to later watch out for

Prioritize accuracy 



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

COMSYS | RWTH AACHEN UNIVERSITY

# ASSESS – Identifying Pitfalls in Contributions

- **Consider 8 pitfalls from ARP et al.**
  - Sampling bias
  - Label inaccuracy
  - Data snooping
  - Spurious correlations
  - Biased parameters
  - Inappropriate baselines
  - Inappropriate measures
  - Lab-only evaluation
  - Inappropriate threat model
  - Base rate fallacy
- **Same technical approach as for RATE**
  - Find concept vector per pitfall
  - Project hidden state onto concept vector

Detect flaws to later watch out for

Prioritize accuracy

Independent of the model, only require access to final layer

>90% accuracy for most pitfalls

Llama-3.2-3B-Instruct



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Combining the Stages



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages






Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

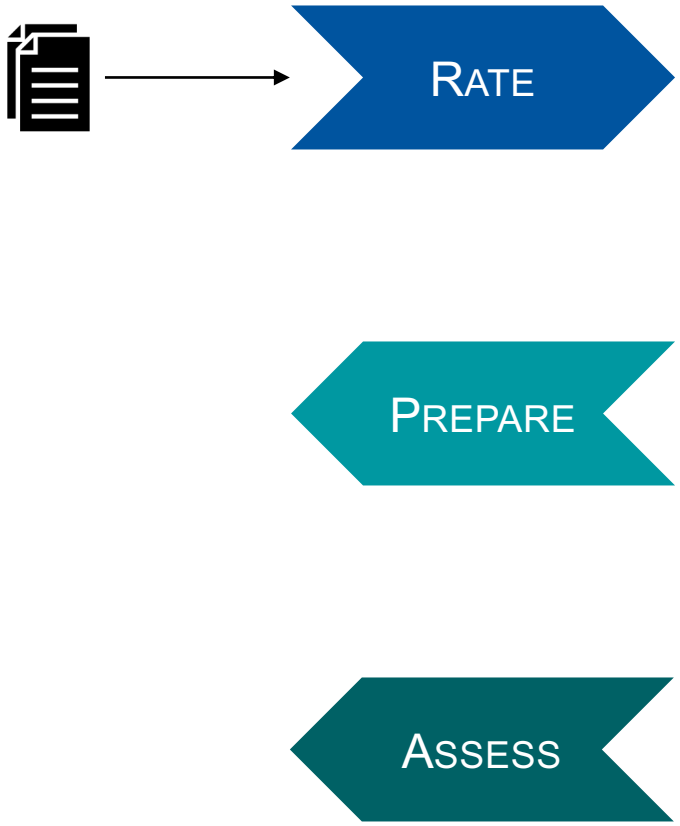



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

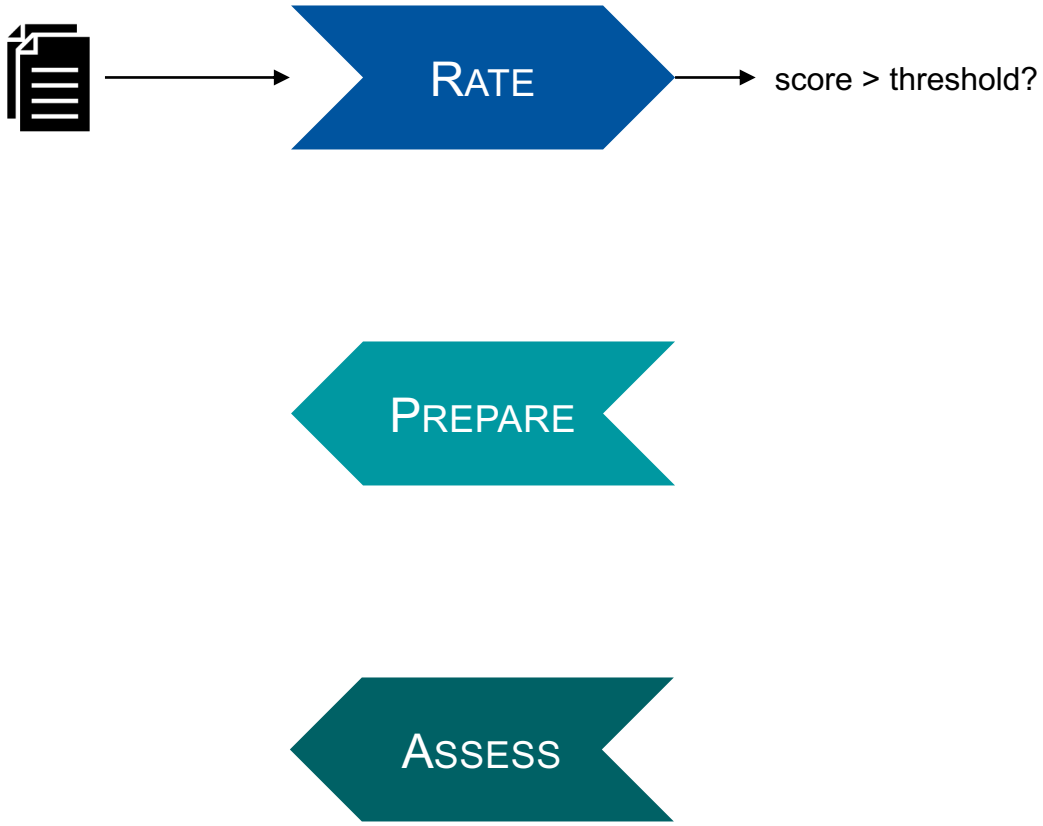



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

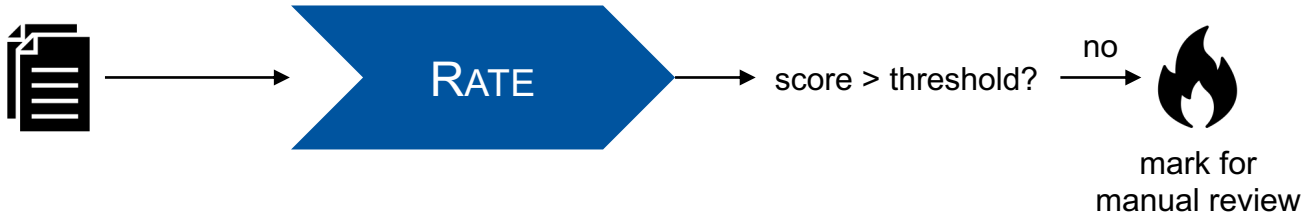



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages




PREPARE

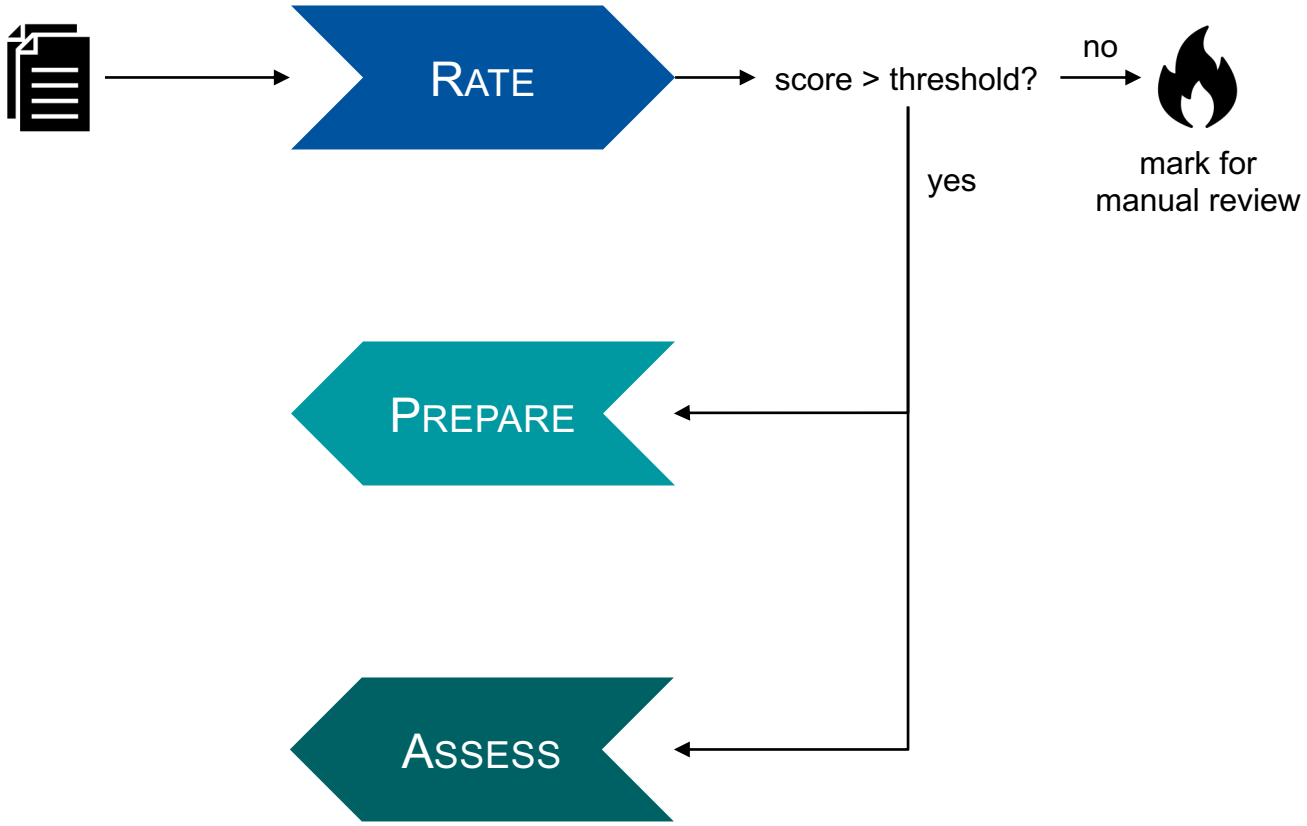
ASSESS

Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

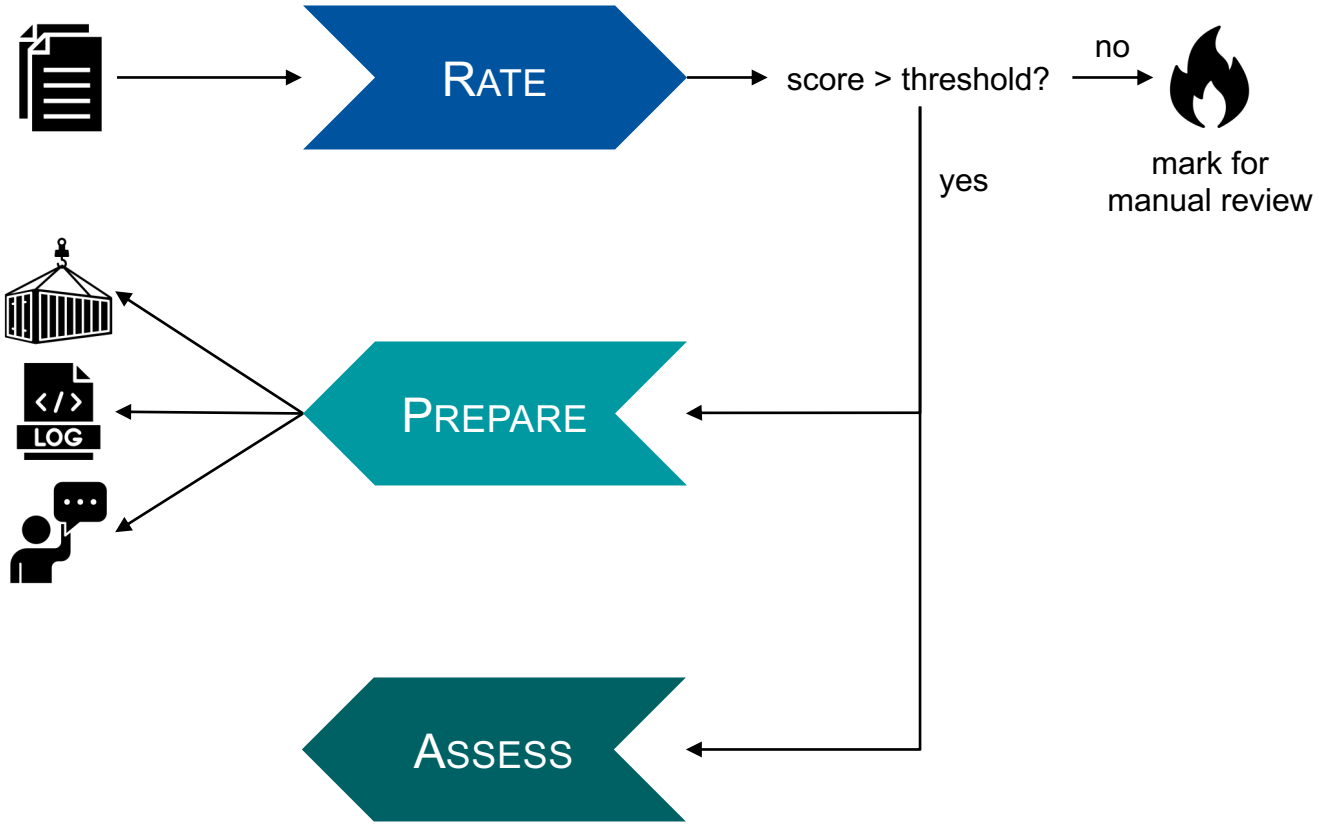



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

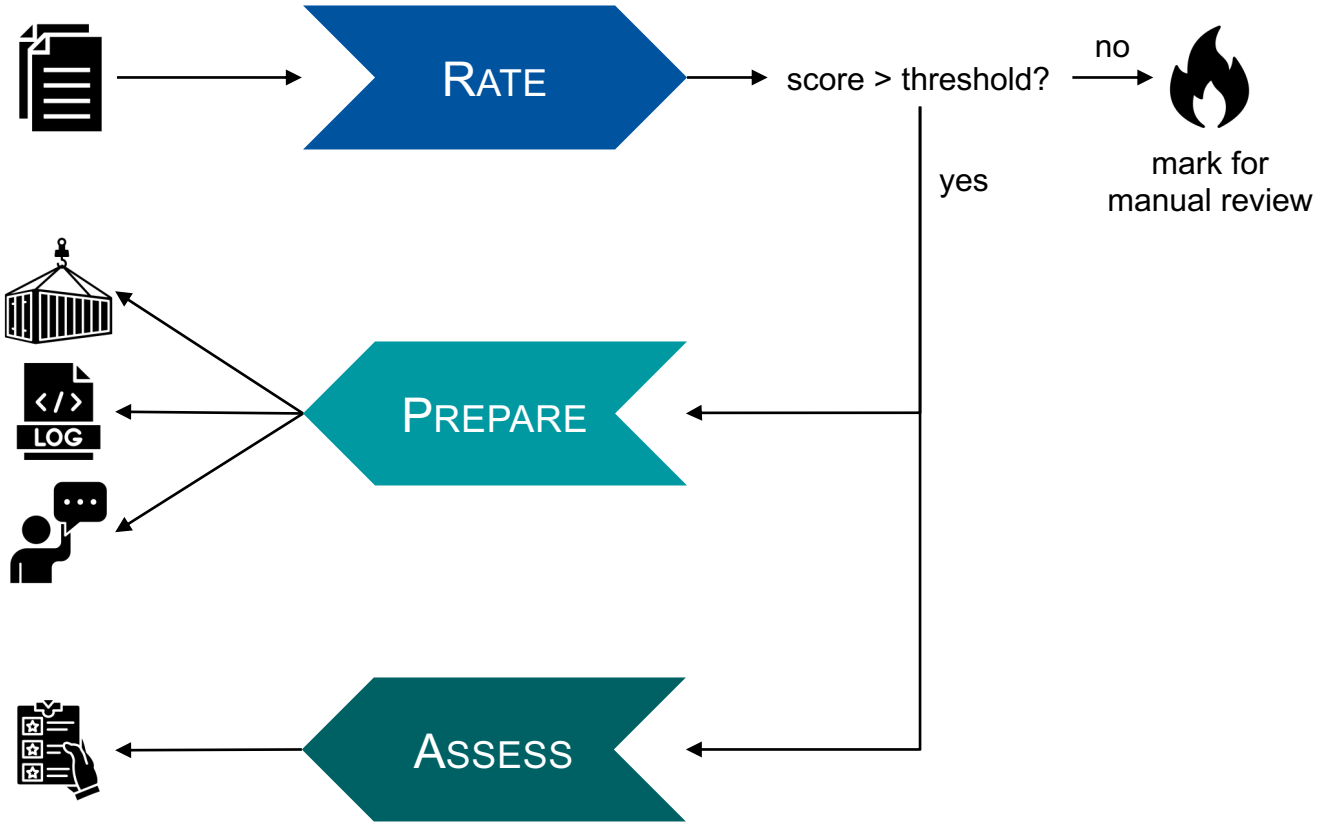



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

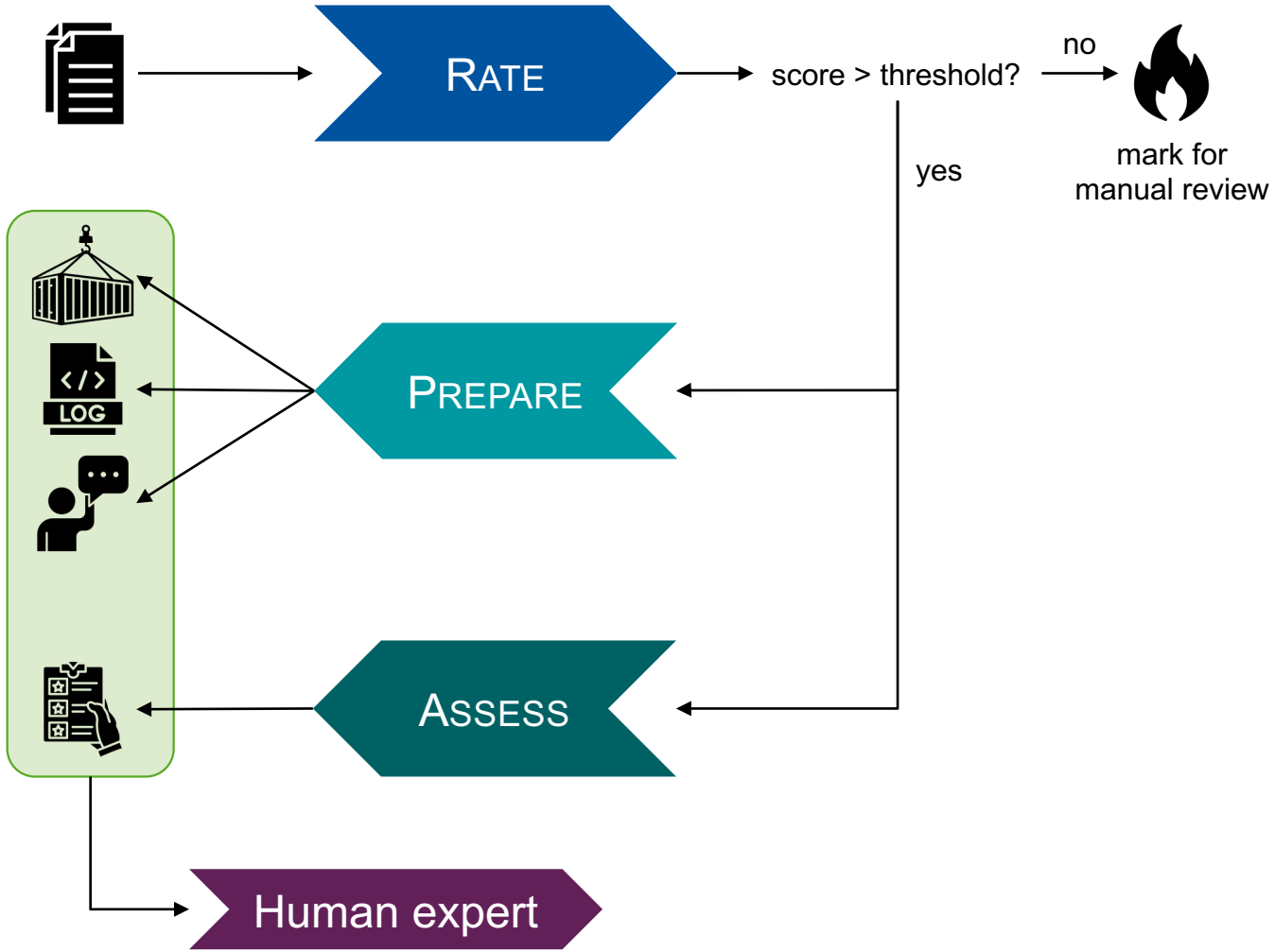



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

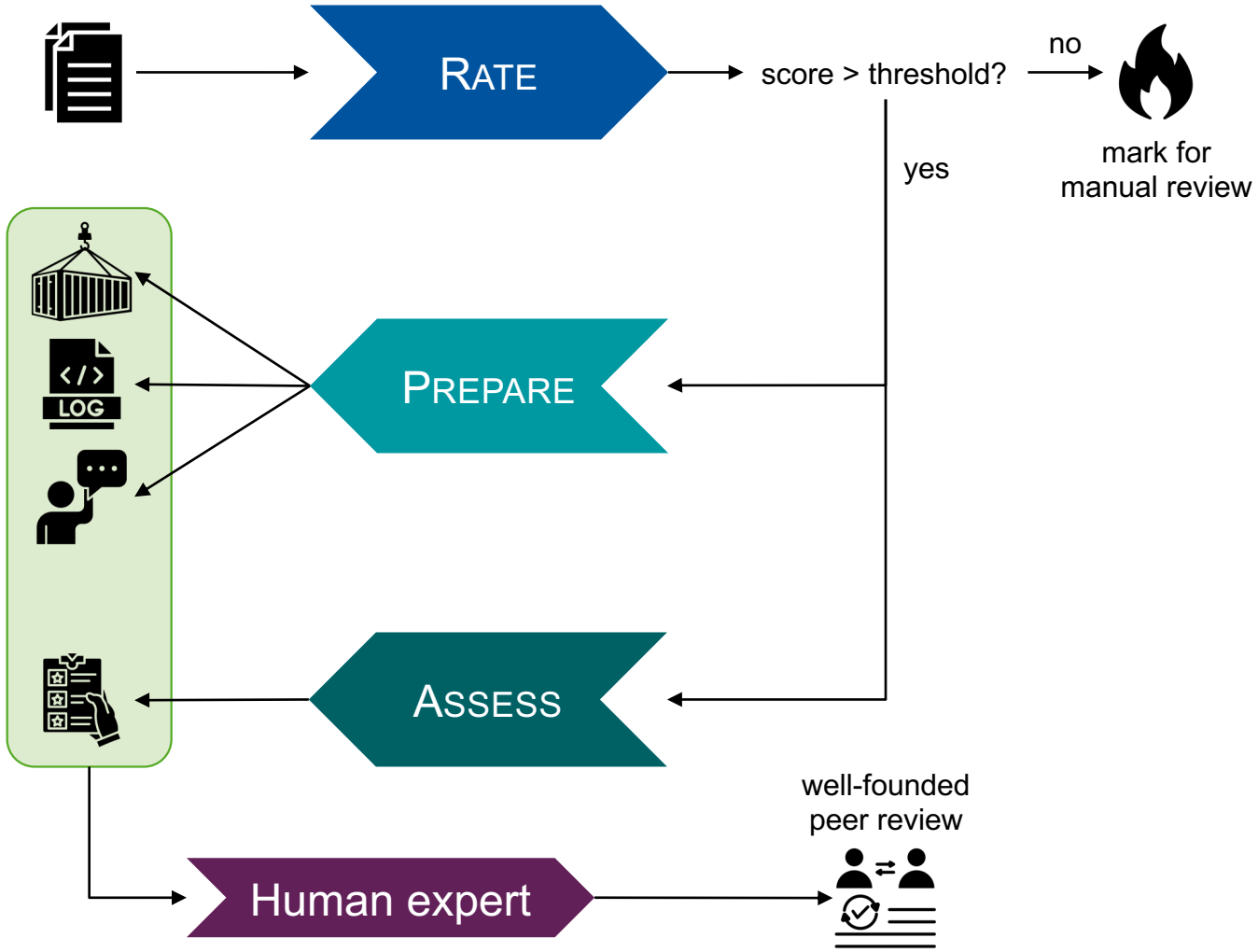



Support the human artifact evaluator



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Combining the Stages

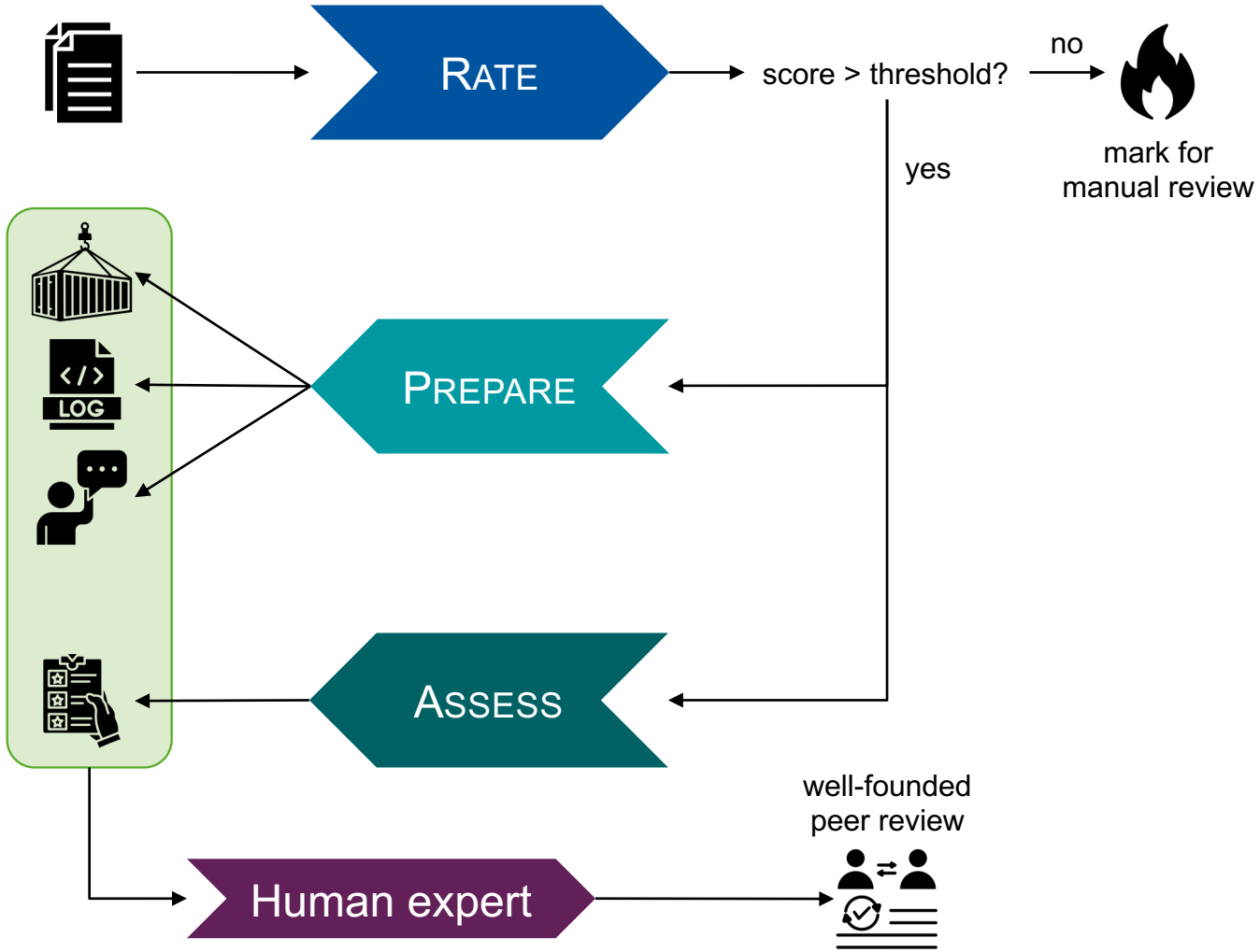



Support the human artifact evaluator



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Combining the Stages



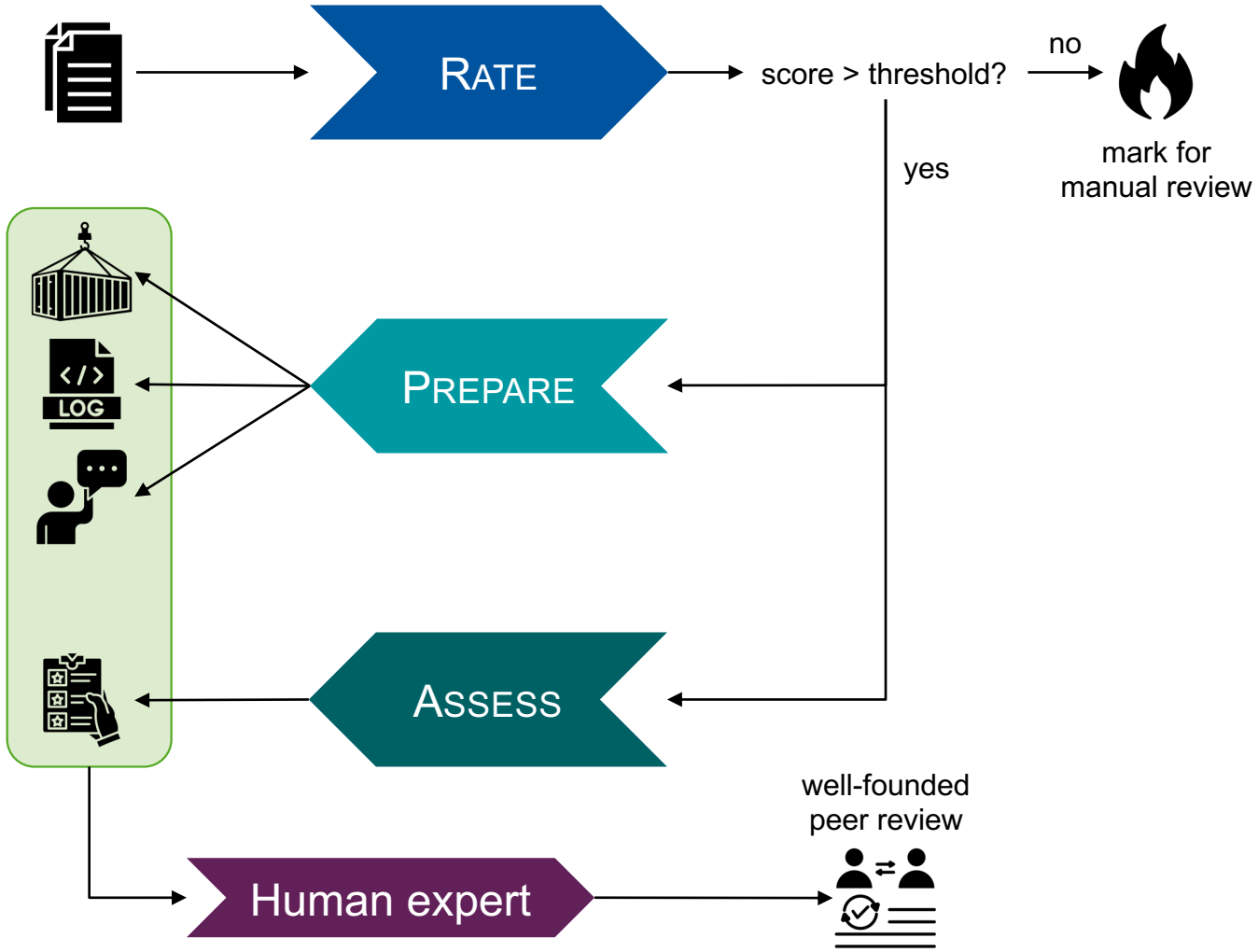

Support the human artifact evaluator

Maximize accuracy



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Combining the Stages




- Support the human artifact evaluator
- Maximize accuracy
- Expert gets ready-to-use execution environments
- 28% successful code setup
  85% true negative rate
  72% accuracy



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Experimental Setup and Datasets



- **LLM considerations**
  - Focus on a locally-hosted, open-source model for privacy reasons
    - This approach also reduces the “direct” costs (e.g., API subscriptions)
    - LLAMA-3.2-3B-INSTRUCT running on an NVIDIA H-100 Tensor-Core GPU



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Experimental Setup and Datasets

- **LLM considerations**
  - Focus on a locally-hosted, open-source model for privacy reasons
    - This approach also reduces the “direct” costs (e.g., API subscriptions)
    - LLAMA-3.2-3B-INSTRUCT running on an NVIDIA H-100 Tensor-Core GPU
  - Since this model performed poorly in the PREPARE stage, we tried a commercial one
    - OpenAI’s GPT-4O-MINI
    - The other model generated many invalid commands, impairing performance


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Experimental Setup and Datasets



- **LLM considerations**
  - Focus on a locally-hosted, open-source model for privacy reasons
    - This approach also reduces the “direct” costs (e.g., API subscriptions)
    - LLAMA-3.2-3B-INSTRUCT running on an NVIDIA H-100 Tensor-Core GPU
  - Since this model performed poorly in the PREPARE stage, we tried a commercial one
    - OpenAI’s GPT-4O-MINI
    - The other model generated many invalid commands, impairing performance
- **Runtime environment**
  - 2x Intel Xeon Platinum 8160 CPUs and 2x NVIDIA Tesla V-100 GPU


Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Experimental Setup and Datasets



- **LLM considerations**
  - Focus on a locally-hosted, open-source model for privacy reasons
    - This approach also reduces the “direct” costs (e.g., API subscriptions)
    - LLAMA-3.2-3B-INSTRUCT running on an NVIDIA H-100 Tensor-Core GPU
  - Since this model performed poorly in the PREPARE stage, we tried a commercial one
    - OpenAI’s GPT-4O-MINI
    - The other model generated many invalid commands, impairing performance
- **Runtime environment**
  - 2x Intel Xeon Platinum 8160 CPUs and 2x NVIDIA Tesla V-100 GPU
- **Datasets**
  - OLSZEWSKI study (142-311/744 analyzed papers used)
  - ARP study (30/30 analyzed papers used)

**We use a random sample of 10-12 papers to derive the concept vector (RATE and ASSESS stage)**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the Entire Pipeline

| *Total* | | Olszewski study | | |
|---|---|---|---|---|
| *126* | | runs | ¬runs | |
| Pipeline | runs | 7.14% | 8.73% | 15.87% |
| | ¬runs | 19.05% | 65.08% | 84.13% |
| | | 26.19% | 73.81% | |

**Accuracy: 72.22%** Precision: 45.00% Recall: 27.27%



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the Entire Pipeline

| *Total* | | Olszewski study | | |
|---|---|---|---|---|
| *126* | | runs | ¬runs | |
| Pipeline | runs | 7.14% | 8.73% | 15.87% |
| | ¬runs | 19.05% | 65.08% | 84.13% |
| | | 26.19% | 73.81% | |

**Accuracy: 72.22%** Precision: 45.00% Recall: 27.27%

**1st:** Prioritize on correct decisions for as many submissions as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the Entire Pipeline

| *Total* | | Olszewski study | | |
|---|---|---|---|---|
| *126* | | runs | ¬runs | |
| Pipeline | runs | 7.14% | 8.73% | 15.87% |
| | ¬runs | 19.05% | 65.08% | 84.13% |
| | | 26.19% | 73.81% | |

**Accuracy: 72.22%** Precision: 45.00% Recall: 27.27%

True negative rate of over 88%

**1st:** Prioritize on correct decisions for as many submissions as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the Entire Pipeline

| *Total* | | Olszewski study | | |
|---|---|---|---|---|
| *126* | | runs | ¬runs | |
| Pipeline | runs | 7.14% | 8.73% | 15.87% |
| | ¬runs | 19.05% | 65.08% | 84.13% |
| | | 26.19% | 73.81% | |

**Accuracy: 72.22%** Precision: 45.00% Recall: 27.27%

True negative rate of over 88%

**1st:** Prioritize on correct decisions for as many submissions as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Low recall due to limitations of the evaluation environment (lack of GUI and hardware support)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the Entire Pipeline

| *Total* 126 | | Olszewski study: runs | Olszewski study: ¬runs | |
|---|---|---|---|---|
| Pipeline | runs | 7.14% | 8.73% | 15.87% |
| | ¬runs | 19.05% | 65.08% | 84.13% |
| | | 26.19% | 73.81% | |

**Accuracy: 72.22%** Precision: 45.00% Recall: 27.27%

True negative rate of over 88%

**1st:** Prioritize on correct decisions for as many submissions as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Low recall due to limitations of the evaluation environment (lack of GUI and hardware support)

Future research opportunities



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the RATE Stage

| *Total* | | OLSZEWSKI study | | |
|---|---|---|---|---|
| *130* | | runs | ¬runs | |
| RATE | runs | 40.77% | 54.62% | 95.38% |
| | ¬runs | 2.31% | 2.31% | 84.13% |
| | | 43.08% | 56.92% | |

Accuracy: 43.08% Precision: 42.74% **Recall: 94.64%**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the RATE Stage

| *Total* | | OLSZEWSKI study | | |
|---|---|---|---|---|
| *130* | | runs | ¬runs | |
| RATE | runs | 40.77% | 54.62% | 95.38% |
| | ¬runs | 2.31% | 2.31% | 84.13% |
| | | 43.08% | 56.92% | |

Accuracy: 43.08% Precision: 42.74% **Recall: 94.64%**

**1st:** Prioritize on high recall to not wrongfully discard reproducible papers
**2nd:** Minimize number of false positives to avoid overburdening subsequent stages



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the RATE Stage

| *Total* | | OLSZEWSKI study | | |
|---|---|---|---|---|
| *130* | | runs | ¬runs | |
| RATE | runs | 40.77% | 54.62% | 95.38% |
| | ¬runs | 2.31% | 2.31% | 84.13% |
| | | 43.08% | 56.92% | |

Accuracy: 43.08% Precision: 42.74% **Recall: 94.64%**

True negative rate is only 4%, improvements needed (!)

**1st:** Prioritize on high recall to not wrongfully discard reproducible papers
**2nd:** Minimize number of false positives to avoid overburdening subsequent stages



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the RATE Stage

| *Total* | | OLSZEWSKI study | | |
|---|---|---|---|---|
| *130* | | runs | ¬runs | |
| RATE | runs | 40.77% | 54.62% | 95.38% |
| | ¬runs | 2.31% | 2.31% | 84.13% |
| | | 43.08% | 56.92% | |

Accuracy: 43.08% Precision: 42.74% **Recall: 94.64%**

True negative rate is only 4%, improvements needed (!)

But: Explainability of the “abstracted” concept vector is missing

**1st:** Prioritize on high recall to not wrongfully discard reproducible papers
**2nd:** Minimize number of false positives to avoid overburdening subsequent stages



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the PREPARE Stage

| *Total* *311* | | OLSZEWSKI study | | |
|---|---|---|---|---|
| | | runs | ¬runs | |
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the PREPARE Stage

| *Total* 311 | | OLSZEWSKI study | | |
|---|---|---|---|---|
| | | runs | ¬runs | |
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%

**1st:** Prioritize accuracy since we want to make correct decisions for as many papers as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Detailed Performance of the PREPARE Stage

| *Total* 311 | | OLSZEWSKI study runs | OLSZEWSKI study ¬runs | |
|---|---|---|---|---|
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%

**1st:** Prioritize accuracy since we want to make correct decisions for as many papers as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Again, issues related to the evaluation environment (no conceptual flaw in itself)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the PREPARE Stage

| *Total* *311* | | OLSZEWSKI study runs | OLSZEWSKI study ¬runs | |
|---|---|---|---|---|
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

Avoids manual human effort for 1/3

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%

**1st:** Prioritize accuracy since we want to make correct decisions for as many papers as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Again, issues related to the evaluation environment (no conceptual flaw in itself)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the PREPARE Stage

| *Total* *311* | | OLSZEWSKI study runs | OLSZEWSKI study ¬runs | |
|---|---|---|---|---|
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

We even set up three artifacts the OLSZEWSKI study could not (!)

Avoids manual human effort for 1/3

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%

**1st:** Prioritize accuracy since we want to make correct decisions for as many papers as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Again, issues related to the evaluation environment (no conceptual flaw in itself)



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Detailed Performance of the PREPARE Stage

| *Total* *311* | | OLSZEWSKI study runs | OLSZEWSKI study ¬runs | |
|---|---|---|---|---|
| PREPARE | runs | 7.40% | 14.79% | 22.19% |
| | ¬runs | 18.97% | 58.84% | 77.81% |
| | | 26.37% | 73.63% | |

We even set up three artifacts the OLSZEWSKI study could not (!)

Avoids manual human effort for 1/3

**Accuracy: 66.24%** Precision: 33.33% Recall: 28.05%

**1st:** Prioritize accuracy since we want to make correct decisions for as many papers as possible
**2nd:** Maximize number of true positives since false negatives can still be handled manually

Again, issues related to the evaluation environment (no conceptual flaw in itself)

Combining the stages improves overall performance



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Details of the ASSESS Stage



- **Given the low number of evaluated papers (only 30), the evaluation is limited**
  - Particularly, the number of pitfalls present vs. not present varies greatly across pitfalls
  - However, a large dataset would have introduced significant manual effort, beyond our paper's scope



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Details of the ASSESS Stage



- **Given the low number of evaluated papers (only 30), the evaluation is limited**
  - Particularly, the number of pitfalls present vs. not present varies greatly across pitfalls
  - However, a large dataset would have introduced significant manual effort, beyond our paper's scope

- **Further, the dataset is constrained, given its background on machine learning for security**
  - How much can other research directions and domains can profit from the ASSESS stage?
  - Looking at the bigger picture: Integrate the ASSESS stage into LLM-assisted paper review?
    - Future work should investigate respective synergies *(we came from the artifact angle in this work)*



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Details of the ASSESS Stage



- **Given the low number of evaluated papers (only 30), the evaluation is limited**
  - Particularly, the number of pitfalls present vs. not present varies greatly across pitfalls
  - However, a large dataset would have introduced significant manual effort, beyond our paper's scope

- **Further, the dataset is constrained, given its background on machine learning for security**
  - How much can other research directions and domains can profit from the ASSESS stage?
  - Looking at the bigger picture: Integrate the ASSESS stage into LLM-assisted paper review?
    - Future work should investigate respective synergies *(we came from the artifact angle in this work)*

More work needed but the general approach of using concept vectors seems generalizable



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Limitations and Lessons Learned

- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models
- **Datasets**
- **Model tradeoffs**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Limitations and Lessons Learned

- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models
- **Datasets**
- **Model tradeoffs**

**Employ desktop agents, e.g., UI-TARS?**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Limitations and Lessons Learned

- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models
- **Datasets**
  - Small dataset sizes
    - OLSZEWSKI: **<300** artifacts
    - ARP: **30** papers
  - No similar datasets available
- **Model tradeoffs**

**Employ desktop agents, e.g., UI-TARS?**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Limitations and Lessons Learned

- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models

- **Datasets**
  - Small dataset sizes
    - OLSZEWSKI: **<300** artifacts
    - ARP: **30** papers
  - No similar datasets available

- **Model tradeoffs**

**Employ desktop agents, e.g., UI-TARS?**

**Create more datasets! Shadow artifact evaluation?**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Limitations and Lessons Learned

- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models

- **Datasets**
  - Small dataset sizes
    - OLSZEWSKI: **<300** artifacts
    - ARP: **30** papers
  - No similar datasets available

- **Model tradeoffs**
  - Commercial models much faster than local models
    - Privacy favors local deployment
  - Commercial models perform better with source code

**Employ desktop agents, e.g., UI-TARS?**

**Create more datasets! Shadow artifact evaluation?**



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Limitations and Lessons Learned



- **Technical constraints**
  - No GUI environments
    → cannot run, e.g., browsers
  - Only limited GPU access
    → cannot train large models

- **Datasets**
  - Small dataset sizes
    - OLSZEWSKI: **<300** artifacts
    - ARP: **30** papers
  - No similar datasets available

- **Model tradeoffs**
  - Commercial models much faster than local models
    - Privacy favors local deployment
  - Commercial models perform better with source code

**Employ desktop agents, e.g., UI-TARS?**

**Create more datasets! Shadow artifact evaluation?**

**Try different models, and fine-tune LLMs**



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Putting our Work into Perspective



- **This research is work in progress**
  - Our contribution is rather about pursuing the idea and the concept and less about providing actual results



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Putting our Work into Perspective



- **This research is work in progress**
  - Our contribution is rather about pursuing the idea and the concept and less about providing actual results
  - Other models may improve performance and reliability
    - GPT-5 has been released since our evaluation
    - However, privacy and IP considerations are arguments to rely on local, open-source models instead



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Putting our Work into Perspective



- **This research is work in progress**
  - Our contribution is rather about pursuing the idea and the concept and less about providing actual results
  - Other models may improve performance and reliability
    - GPT-5 has been released since our evaluation
    - However, privacy and IP considerations are arguments to rely on local, open-source models instead

- **What happens if we apply our paper?**
  - **RATE**: average score, not particularly “good”
  - **PREPARE**: the prepare stage is able to set up the prepare stage itself ☺
    we can also automatically run our other proposed stages within the prepare stage
  - **ASSESS**: our approaches criticizes the lab-only evaluation and an inappropriate threat model



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Putting our Work into Perspective



- **This research is work in progress**
  - Our contribution is rather about pursuing the idea and the concept and less about providing actual results
  - Other models may improve performance and reliability
    - GPT-5 has been released since our evaluation
    - However, privacy and IP considerations are arguments to rely on local, open-source models instead
- **We are curious to see what future work achieves**
  - Thus, we are also committed to open science



github.com/COMSYS/artifact-evaluation-llm-support

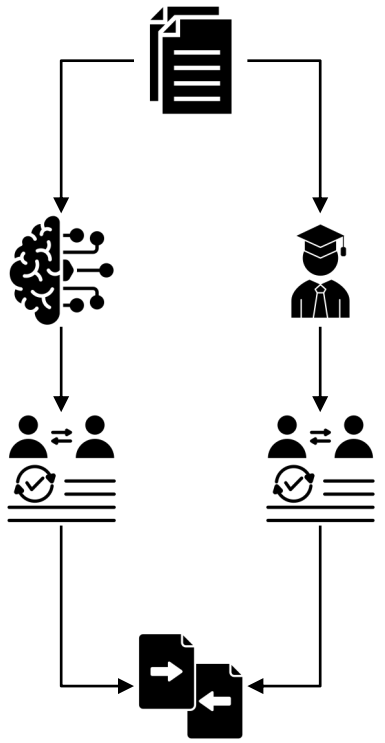
- **What happens if we apply our paper?**
  - **RATE**: average score, not particularly “good”
  - **PREPARE**: the prepare stage is able to set up the prepare stage itself ☺
    we can also automatically run our other proposed stages within the prepare stage
  - **ASSESS**: our approaches criticizes the lab-only evaluation and an inappropriate threat model



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Next Steps: Integration Into Peer Review

- **Proposal: shadow AE at current conferences**
  - Assess performance of the tools, find points to improve → low-risk pilot phase

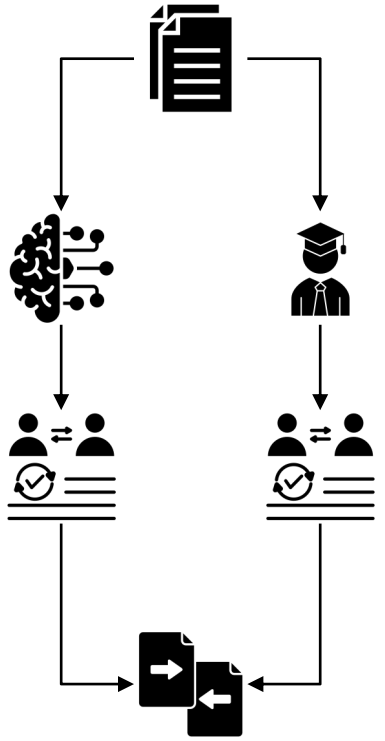




Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Next Steps: Integration Into Peer Review

- **Proposal: shadow AE at current conferences**
  - Assess performance of the tools, find points to improve → low-risk pilot phase
- **Reviewer benefits**
  - Avoid **trial-and-error setups**
  - **Reduce overall workload** on artifact evaluators
  - Structured reports enable **richer feedback**

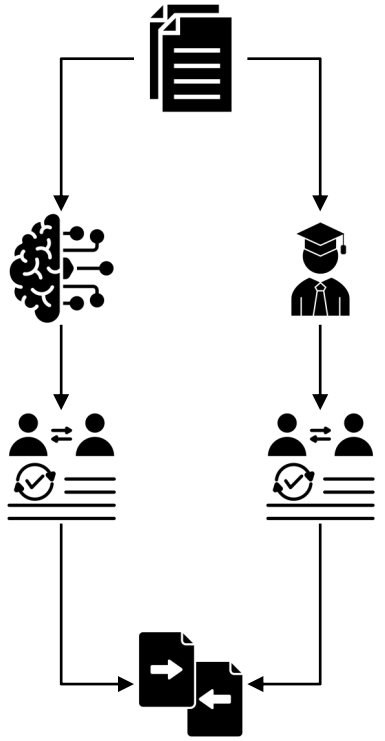




Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Next Steps: Integration Into Peer Review

- **Proposal: shadow AE at current conferences**
  - Assess performance of the tools, find points to improve → low-risk pilot phase
- **Reviewer benefits**
  - Avoid **trial-and-error setups**
  - **Reduce overall workload** on artifact evaluators
  - Structured reports enable **richer feedback**
- **Community impact**
  - **Transparent criteria** and tools
  - Encourage **higher quality** submissions
  - **More consistency** across reviewers and conferences
  - Open-source releases encourage **community adoption** and improvement





Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Acknowledgments


A thanks to those who enabled our research and this presentation!

## Datasets

Olszewski study (2023)

Arp study (2022)

Daniel Arp even actively supported this work!

## Co-Authors

- David
- Robin
- Karl
- Johannes
- Anastasiia
- Sandra
- Klaus



David for preparing parts of this presentation

## Funding Bodies

SUSTAINET guardian

CELTIC-NEXT Eureka Cluster

With funding from the:

Federal Ministry of Research, Technology and Space

**Travel Expenses** GSO* Guidance, Skills & Opportunities for Researchers

Klaus Tschira Stiftung

Thanks for the invitation, David!





Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Conclusion



- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Conclusion



- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)
- **Three-step toolkit** 

  - **RATE**: reproducibility score based on paper, leverage LLM's internal state
  - **PREPARE**: autonomously set up code execution environments
  - **ASSESS**: identify methodological and experimental flaws in submissions



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Conclusion



- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)
- **Three-step toolkit** 

  - **RATE**: reproducibility score based on paper, leverage LLM's internal state
  - **PREPARE**: autonomously set up code execution environments
  - **ASSESS**: identify methodological and experimental flaws in submissions
- **Our work demonstrates the general potential, but more work is needed**
  - Our pipeline currently correctly identifies non-reproducible submissions in **85%** of cases



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**

# Conclusion



- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)
- **Three-step toolkit** 

  - **RATE**: reproducibility score based on paper, leverage LLM's internal state
  - **PREPARE**: autonomously set up code execution environments
  - **ASSESS**: identify methodological and experimental flaws in submissions
- **Our work demonstrates the general potential, but more work is needed**
  - Our pipeline currently correctly identifies non-reproducible submissions in **85%** of cases

**Next Goal:** A community effort for more sustainable artifact evaluation at scale



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

# Conclusion

- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)
- **Three-step toolkit** RATE → PREPARE → ASSESS
  - **RATE**: reproducibility score based on paper, leverage LLM's internal state
  - **PREPARE**: autonomously set up code execution environments
  - **ASSESS**: identify methodological and experimental flaws in submissions



- **Our work demonstrates the general potential, but more work is needed**
  - Our pipeline currently correctly identifies non-reproducible submissions in **85%** of cases

**Next Goal:** A community effort for more sustainable artifact evaluation at scale

David Heye, Karl Kindermann, Robin Decker, Johannes Lohmöller, Anastasiia Belova, Sandra Geisler, Klaus Wehrle, Jan Pennekamp. **Supporting Artifact Evaluation with LLMs: A Study with Published Security Research Papers**. *2025 IEEE International Conference on Big Data.* To be published.

github.com/COMSYS/artifact-evaluation-llm-support



Jan Pennekamp
pennekamp@comsys.rwth-aachen.de

COMSYS | RWTH AACHEN UNIVERSITY

# Conclusion

- **LLM-driven peer-review support may automate core parts of artifact evaluation**
  - Reducing manual workload without sacrificing accuracy or rigor (in times with high submission volumes)
- **Three-step toolkit** RATE → PREPARE → ASSESS
  - **RATE**: reproducibility score based on paper, leverage LLM's internal state
  - **PREPARE**: autonomously set up code execution environments
  - **ASSESS**: identify methodological and experimental flaws in submissions



- **Our work demonstrates the general potential, but more work is needed**
  - Our pipeline currently correctly identifies non-reproducible submissions in **85%** of cases

**Next Goal:** A community effort for more sustainable artifact evaluation at scale

Thank you for your attention!

David Heye, Karl Kindermann, Robin Decker, Johannes Lohmöller, Anastasiia Belova, Sandra Geisler, Klaus Wehrle, Jan Pennekamp. **Supporting Artifact Evaluation with LLMs: A Study with Published Security Research Papers**. *2025 IEEE International Conference on Big Data.* To be published.

github.com/COMSYS/artifact-evaluation-llm-support



Jan Pennekamp
**pennekamp@comsys.rwth-aachen.de**