

ergon

Roman
Wixinger
Data Scientist
(Ergon)

ETH zürich

Hannes
Stählin
Data Scientist
(Ergon, ETH)

**From demos to
dependability:
Practical
evaluation
strategies for
AI applications**

RSECon2025



From demos to dependability: Practical evaluation strategies for AI applications

1. Introduction
2. Evaluation Strategies
3. Case Studies
4. Conclusion

Introduction: What kind of AI systems do RSEs usually build?

Applied applications

AI systems that are used to improve the operations of the research institute

- Deep research tool for the university library
- Course chatbot
- Automated grading of multiple-choice tests
- Use LLMs to reduce ambiguity in exams questions

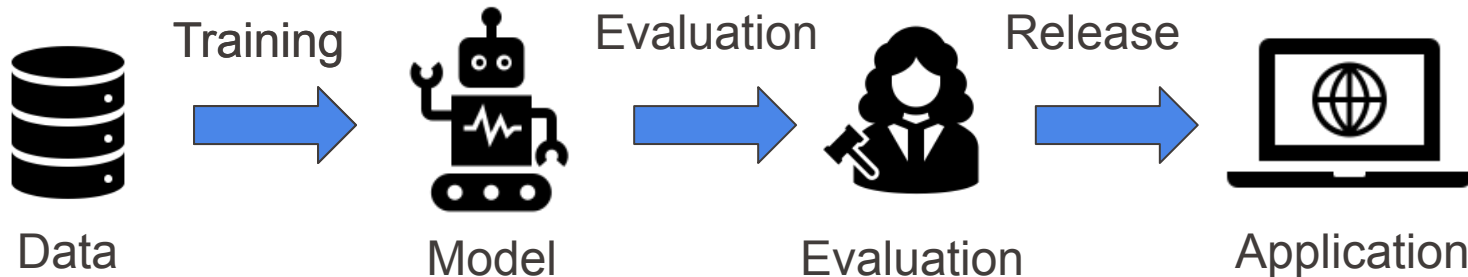
Research applications

AI systems that are part of the research itself, either as subject of interest or as a tool

- Computer Vision models for cell classification
- Legal decision making with LLMs
- Efficient quantization methods for foundation models

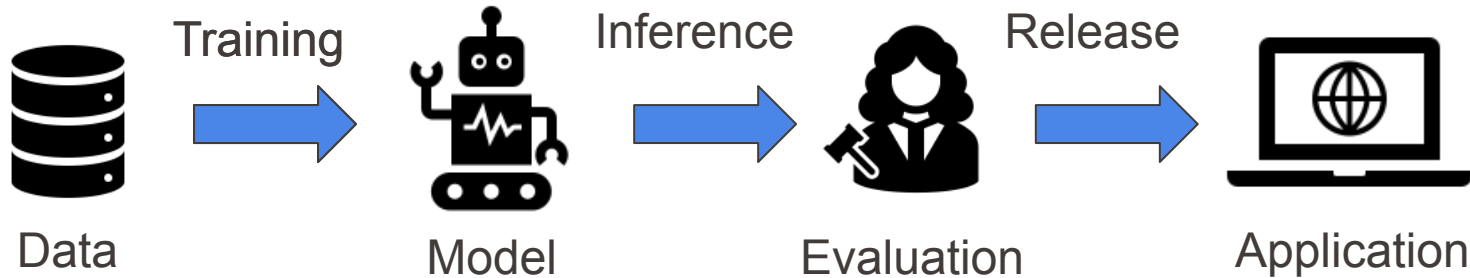
Introduction: How are these AI systems usually developed?

Back in the days (Machine Learning)



Introduction: How are these AI systems usually developed?

Back in the day (Machine Learning)



Classify cats and dogs:

```
{“image_url”: “...”, “cat”,  
“Image_url”: “...”, “dog”}
```

Evaluation metrics

- Accuracy
- Recall
- F1-Score
- Confusion matrix

Introduction: But today we have foundation models

Definition of the term “Foundation model”

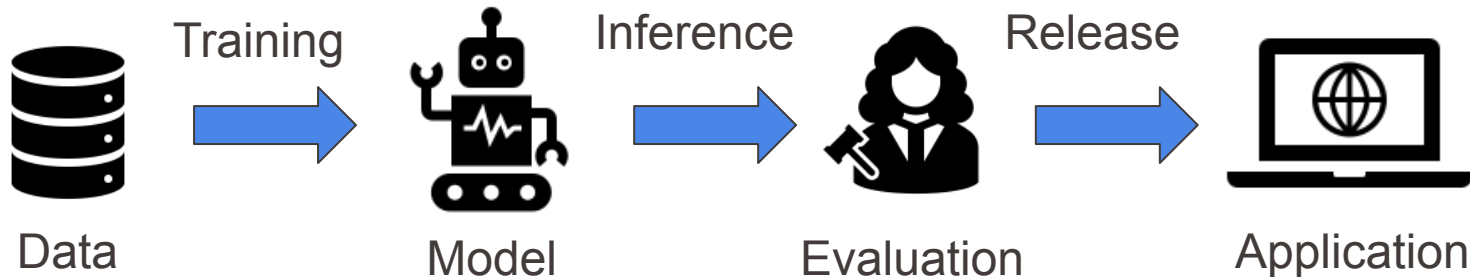
AI model that

- is trained on broad data at scale
- is designed for generality of outputs
- and can be adapted to a wide range of distinctive tasks

Bommasani et al. (2021), On the Opportunities and Risks of Foundation Models.

Introduction: How are these AI systems usually developed?

Back in the day (Machine Learning)



Many people today



Introduction: We need to bring back data and evaluation to AI

Evaluation is crucial for

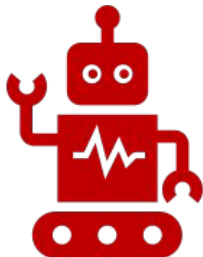
- Making informed decisions about the use of AI
- Build trust with stakeholders
- Optimize performance
- Produce reliable findings and systems

... and a key part of the scientific method!

From demos to dependability: Practical evaluation strategies for AI applications

1. Introduction
- 2. Evaluation Strategies**
3. Case Studies
4. Conclusion

Evaluation strategies



Manual testing



Hand-crafted datasets



Synthetic datasets

Example: RAG Demo on Federal Court Rulings (2023)



BGE Assistant

Wann gibt es eine Verletzung des Beschleunigungsgebots bei der maximalen Frist der 96 stunden Regel nach Art 224 und 226 StPO?

Suchen

Score	ID
80	1B_174/2023
70	1B_608/2011

Der Artikel behandelt die Frage, wann es eine Verletzung des Beschleunigungsgebots bei der maximalen Frist der 96-stunden Regel nach Art 224 und 226 StPO gibt. Genauer gesagt wird die Frage anhand des konkreten Falls beurteilt, ob die Nichteinhaltung der 96-Stunden-Frist durch das Zwangsmassnahmengericht eine Verletzung des Beschleunigungsgebots in Haftsachen darstellt (Absatz 2.2 und 2.3 des Artikels). Die Antwort des Bundesgerichts ist negativ, da die Fristüberschreitung durch das Verhalten der Verteidigung verursacht wurde und im Interesse der Beschwerdeführerin lag. Die Information findet sich im zweiten Erwägungsteil des Artikels.

Sachverhalt:

A.

Die Staatsanwaltschaft Winterthur/Unterland führt eine Strafuntersuchung gegen A._____ wegen rechtswidriger Einreise, rechtswidrigen Aufenthalts und Erwerbstätigkeit ohne Bewilligung. Sie wurde am 2. Februar 2023 festgenommen. Mit Verfügung vom 6. Februar 2023 ordnete das Bezirksgericht Dielsdorf als zuständiges Zwangs-massnahmengericht für A._____ Untersuchungshaft bis vorläufig längstens am 6. Mai 2023 an. Die dagegen erhobene Beschwerde hiess das Obergericht des Kantons Zürich mit Beschluss vom 15. März 2023 teilweise gut und hielt im Urteilsdispositiv fest, dass der Haftantrag der Staatsanwaltschaft vom 4. Februar 2023 formfehlerhaft eingereicht wurde. Soweit die gerichtliche Feststellung einer Verletzung des besonderen Beschleunigungsgebots in Haftsachen beantragt wurde, hiess es die Beschwerde ab.

B.

Mit Beschwerde in Strafsachen vom 29. März 2023 beantragt A._____, der Beschluss des Obergerichts vom 15. März 2023 sei aufzuheben, soweit er nicht die Feststellung der Formfehlerhaftigkeit des Haftantrags betrifft. Es sei eine Verletzung des Beschleunigungsgebots in Haftsachen festzustellen. Zudem sei festzustellen, dass der Freiheitsentzug der Beschwerdeführerin vom 6. Februar 2023 von 14:50 Uhr bis 16:00 Uhr rechtswidrig

Example: RAG Demo on Federal Court Rulings (2023)



BGE Assistant

Wann gibt es eine Verletzung des Beschleunigungsgebots bei der maximalen Frist der 96 stunden Regel nach Art 224 und 226 stpo?

Score	ID
80	1B_174/2023
70	1B_608/2011

Der Artikel behandelt die Frage, wann es eine Verletzung des Beschleunigungsgebots nach Art 224 und 226 StPO gibt. Genauer gesagt wird die Frage anhand der Fragestellung, ob die Frist durch das Zwangsmassnahmengericht eine Verletzung des Beschleunigungsgebots darstellt (Art. 224 Abs. 1 StPO). Die Antwort des Bundesgerichts ist negativ, da die Fristüberschreitung im Interesse der Beschwerdeführerin lag. Die Information findet sich im Urteil des Bundesgerichts vom 15. März 2023 (1B_174/2023).

Sachverhalt:

A.

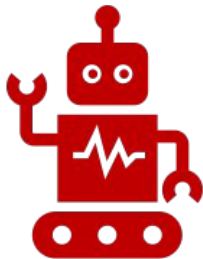
Die Staatsanwaltschaft Winterthur/Unterland führt eine Strafuntersuchung gegen A. (Name) und Erwerbstätigkeit ohne Bewilligung. Sie wurde am 2. Februar 2023 festgenommen. Das Obergericht des Kantons Zürich (Dielsdorf) als zuständiges Zwangs-massnahmengericht für A. (Name) Unterland. Die Beschwerde hiesse das Obergericht des Kantons Zürich mit Beschluss vom 15. März 2023 teilweise gut und hielt im Urteilsdispositiv fest, dass der Haftantrag der Staatsanwaltschaft vom 4. Februar 2023 formfehlerhaft eingereicht wurde. Soweit die gerichtliche Feststellung einer Verletzung des besonderen Beschleunigungsgebots in Haftsachen beantragt wurde, hiess es die Beschwerde ab.

B.

Mit Beschwerde in Strafsachen vom 29. März 2023 beantragt A. (Name), der Beschluss des Obergerichts vom 15. März 2023 sei aufzuheben, soweit er nicht die Feststellung der Formfehlerhaftigkeit des Haftantrags betrifft. Es sei eine Verletzung des Beschleunigungsgebots in Haftsachen festzustellen. Zudem sei festzustellen, dass der Freiheitsentzug der Beschwerdeführerin vom 6. Februar 2023 von 14:50 Uhr bis 16:00 Uhr rechtswidrig war.

Fail: Lawyers search precisely, but we tried vector search for the demo, which is fuzzy.

Evaluation strategies



Manual testing



Hand-crafted datasets



Synthetic datasets

Great for fast, initial iterations, but does not really scale!

Example: Search tool with user feedback

The screenshot shows a search tool interface. At the top, there is a search bar and several filters: 'Alle Bereiche', 'Alle Dokumententypen', and 'Hybride Suche mit Reranking'. Below the search bar is a table with columns: 'Bezeichnung', 'Bereich', 'Typ', 'ID', 'Version', 'PageRank', 'Beste Resultate', and 'Highlight'. The 'Beste Resultate' column contains a list of 'Markieren' buttons, each preceded by a star icon. To the right of the table is a 'Highlight' column. On the far right, there is a 'Feedback' section. It includes an 'Information' box with the following text: 'Willkommen', 'Aktuell sind Dokumente bis zum 3. Juni indiziert.', 'Passendes Dokument gefunden? Stern antippen.', 'Kein Treffer oder allgemeines Feedback? Feld unten ausfüllen.', and 'Vielen Dank für Ihre Rückmeldung!'. Below the information box is a 'Feedback' section with three smiley face icons (happy, neutral, sad) and a text input field with a submit button.

Data Labeling

Example: Search tool with user feedback

Enables automatic
computation of metrics
like Top-k Hit Rate

The screenshot shows a search tool interface. At the top, there is a search bar and several filters: 'Alle Bereiche', 'Alle Dokumententypen', and 'Hybride Suche mit Reranking'. Below the search bar is a table with columns: 'Bezeichnung', 'Bereich', 'Typ', 'ID', 'Version', 'PageRank', 'Beste Resultate', and 'Highlight'. The 'Beste Resultate' column contains a list of 'Markieren' buttons, each preceded by a star icon. To the right of the table is a 'Feedback' section. It includes an 'Information' box with text about document indexing and a feedback prompt. Below this is a 'Feedback' box with three smiley face icons (happy, neutral, sad) and a text input field with a submit button.

Bezeichnung	Bereich	Typ	ID	Version	PageRank	Beste Resultate	Highlight
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	
						★ Markieren	

Information

Willkommen

🔍 Aktuell sind Dokumente bis zum 3. Juni indiziert.

✓ Passendes Dokument gefunden? Stern antippen.

☐ Kein Treffer oder allgemeines Feedback? Feld unten ausfüllen.

Vielen Dank für Ihre Rückmeldung!

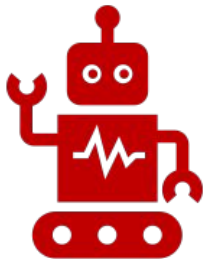
Feedback

😊 😐 😞

➤

Data Labeling

Evaluation strategies



Manual testing



Hand-crafted datasets



Synthetic datasets

Make it easy for users to gather data, which allows for repeated evaluation.

What is synthetic data?

Answer: Synthetic data is data that was created artificially

Examples:



LLM-generated text
or conversation



Images and videos
from game engines



Images created from
geometric shapes



Motivational example: Running LLMs for structured output on smartphones

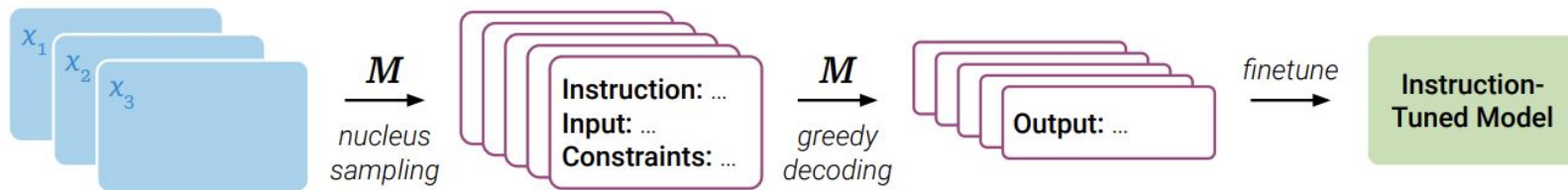
What if we want to use small LLMs on the smartphone?

- **Problem:** Base model does not generate structured output (JSON) well enough
- **Idea:** Fine-tune the model on synthetic data to make it follow the output format
- **Result:** For specialized, domain-specific tasks, we can match the performance of larger models, producing correct JSON output

Example: Unnatural instructions

What to do if human labels are expensive?

➡ Augment your existing data



Honovich, O., Scialom, T., Levy, O., & Schick, T. (2022). Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. *arXiv preprint arXiv:2212.09689*.

Applications in industry

Optimize ML models for domain specific applications, for example execution on the smartphone

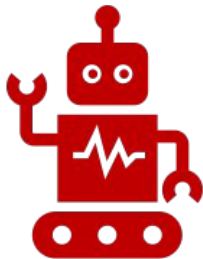


Speech recognition for
specific instructions



Automated document
processing

Evaluation strategies



Manual testing



Hand-crafted datasets

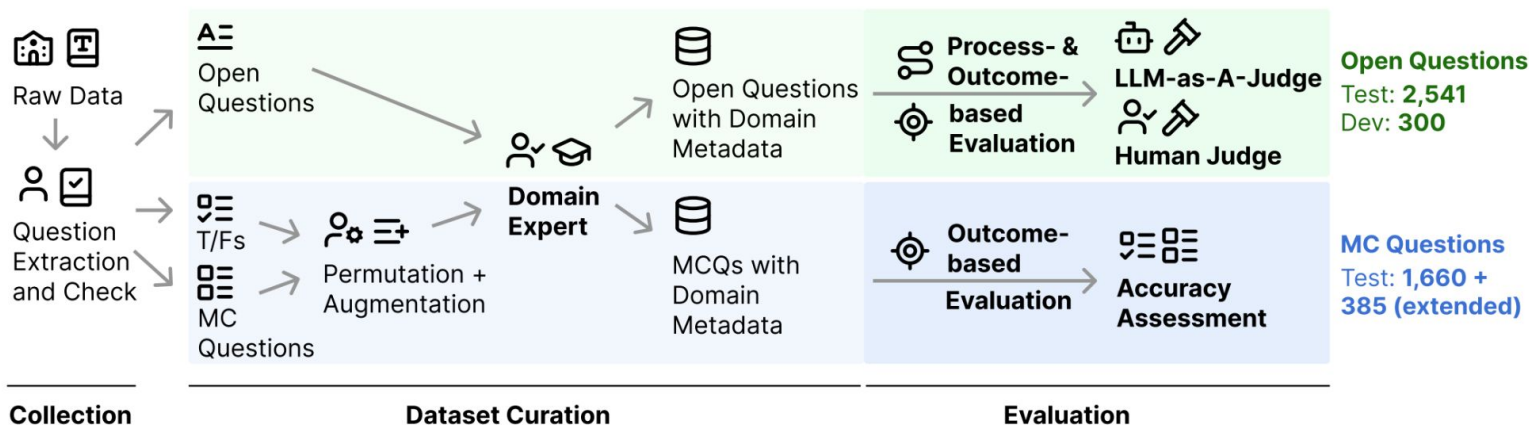


Synthetic datasets

Allows for the rapid generation of massive, fully labeled datasets. But might not represent the real world.

Application of LLMs in research

How can we evaluate legal reasoning with LLMs?



Y. Fan et al. (2025), LEXAM: Benchmarking Legal Reasoning on 340 Law Exams

Example: Alternative Annotator Test

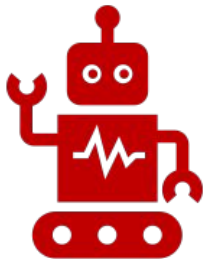
How can we know that using an LLM judge is valid?



Compare alignment with humans

Calderon, N., Reichart, R., & Dror, R. (2025). The Alternative Annotator Test for LLM-as-a-Judge

Evaluation strategies



Manual testing

Great for fast, initial iterations, but does not really scale!



Hand-crafted datasets

Make it easy for users to gather data, which allows for repeated evaluation.



Synthetic datasets

Allows for the rapid generation of massive, fully labeled datasets. But might not represent the real world.

Case Studies

1. **LexCraft (University of Bern)**
2. Precision OncoAssist (University Children's Hospital Zurich)

LexCraft: Digital Decision Support System in Specialized Legal Domains

u^b

b
**UNIVERSITÄT
BERN**

Raphael Zemp
Muhammad Akash Bin Nasir

ergon

Roman Wixinger
Patrick Humbel

1. LexCraft (University of Bern)
2. **Precision OncoAssist (University Children's Hospital Zurich)**

Precision OncoAssist

Personalizing Leukemia Therapy with AI

Healthcare & LifeSciences 2025 Community Kick-off
March 6, 2025

What can and cannot be done with these evaluation strategies

- Manual testing:
 - Great for fast, initial iterations, but does not really scale.
- Hand-crafted datasets:
 - Time-consuming to create and maintain. For example, labeling images or transcribing audio is extremely expensive.
- Synthetic datasets:
 - May not fully capture real-world complexities. The model might perform well on the clean, synthetic data but fail on real-world inputs.
 - Ground methods like LLM as a Judge with methods like the Alternative Annotator Test.

From demos to dependability: Practical evaluation strategies for AI applications

1. Introduction
2. Evaluation Strategies
3. Case Studies
4. **Conclusion**

Checklist for your next AI project

Don'ts

- Start with the method instead of the research question or problem
- Create a demo and just test it yourself – looks good to me

Do's:

- Define clear metrics and KPIs to evaluate your AI model
- Implement reproducible evaluation pipelines from the beginning
- Use software engineering best practices to create reproducible and maintainable model deployments

Key takeaways

Treating evaluation as a first class citizen, you gain:

- Trust
 - Evaluation builds confidence with stakeholders.
- Performance
 - Evaluation drives meaningful optimization.
- Reliability
 - Evaluation ensures robust and reproducible systems.

Testing and evaluation is a key part of good (research) software engineering!



ergon

Roman
Wixinger
Data Scientist
(Ergon)

ETH zürich

Hannes
Stählin
Data Scientist
(Ergon, ETH)

**From demos to
dependability:
Practical
evaluation
strategies for
AI applications**

**Thank
you!**

RSECon2025