



# 20 Years of Data: Building finetuning datasets and fixing search along the way

Niklaus Hofer  
**Uphill Conf**  
8<sup>th</sup> of May 2026

---

# Overview



- About us
- Introduction
- Training pipeline
  - Identifying data sources
  - Extracting data
  - Statistical annotation
  - Finetuning
  - Evaluation
- Conclusion



# About us

# About stepping stone AG and /me



- Cloud provider
- Own infrastructure
- Cloud services
  - IaaS, CaaS, PaaS and SaaS
- Managed services
- Artificial intelligence
  - GPUs
  - AI on Demand
  - Agentic AI
- Niklaus Hofer
- Joined in 2016
- CTO





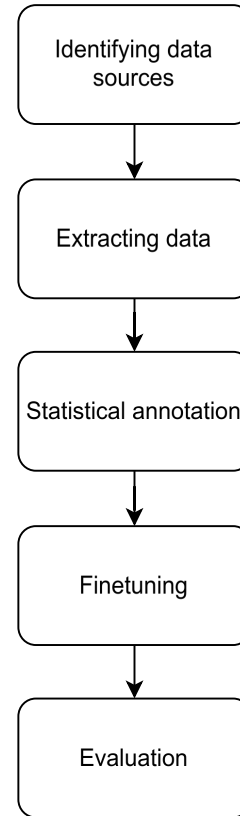
# Introduction

# Serving data to AI



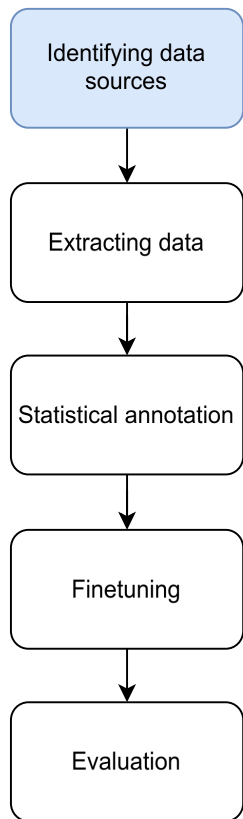
- Load into context
- Access to search
- Retrieval-Augmented Generation (RAG)
- Finetune an existing LLM

# Pipeline





# Identifying data sources



# Data sources




- Documentation
- Ticketing system
- Support e-mail
- Document Management System (DMS)
- Customer Relation Management (CRM)

# Example: MediaWiki



stepping stone



Main page  
Important resources  
Recent changes  
All pages  
All categories  
All files  
Interwikis  
Version  
Help

Categories

- ▶ Accounting
- ▶ Apprentice meetings
- ▶ Apprenticeship
- ▶ Assets
- ▶ Billing
- ▶ Company
- ▶ Corporate design
- ▶ Customers
- ▶ Data centre
- ▶ Documentation
- ▶ Events
- ▶ Hardware
- ▶ Human resources
- ▶ ISMS
- ▶ Maintenance

Sst-nho Talk Preferences Watchlist Contributions Log out

Main Page **Discussion** Read Edit <Visual Editor> More Search openstack

## Main Page

**Emergency numbers** (contact with authorities).  
**Contact numbers** related to the office at Wasserwerksgasse 7.

**Information security management system (ISMS)** [ Edit | <Visual Editor> ]

The category ISMS collects all related documentation.

Information security strategy
Information security strategy
ISMS scope statement
Statement of applicability (SOA)

Information security controls reference
5 Organizational controls
6 People controls
7 Physical controls
8 Technological controls

Information security procedures
Information security procedures
5 Organizational controls procedures
6 People controls procedures

Create a new page / category [ Edit ]  
<Visual Editor> ]

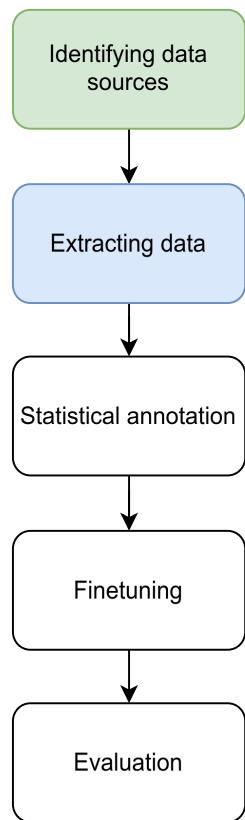
Create a blank page.

Create blank page

Check the Customer VM Naming Convention manual before creating a OpenStack VM page.

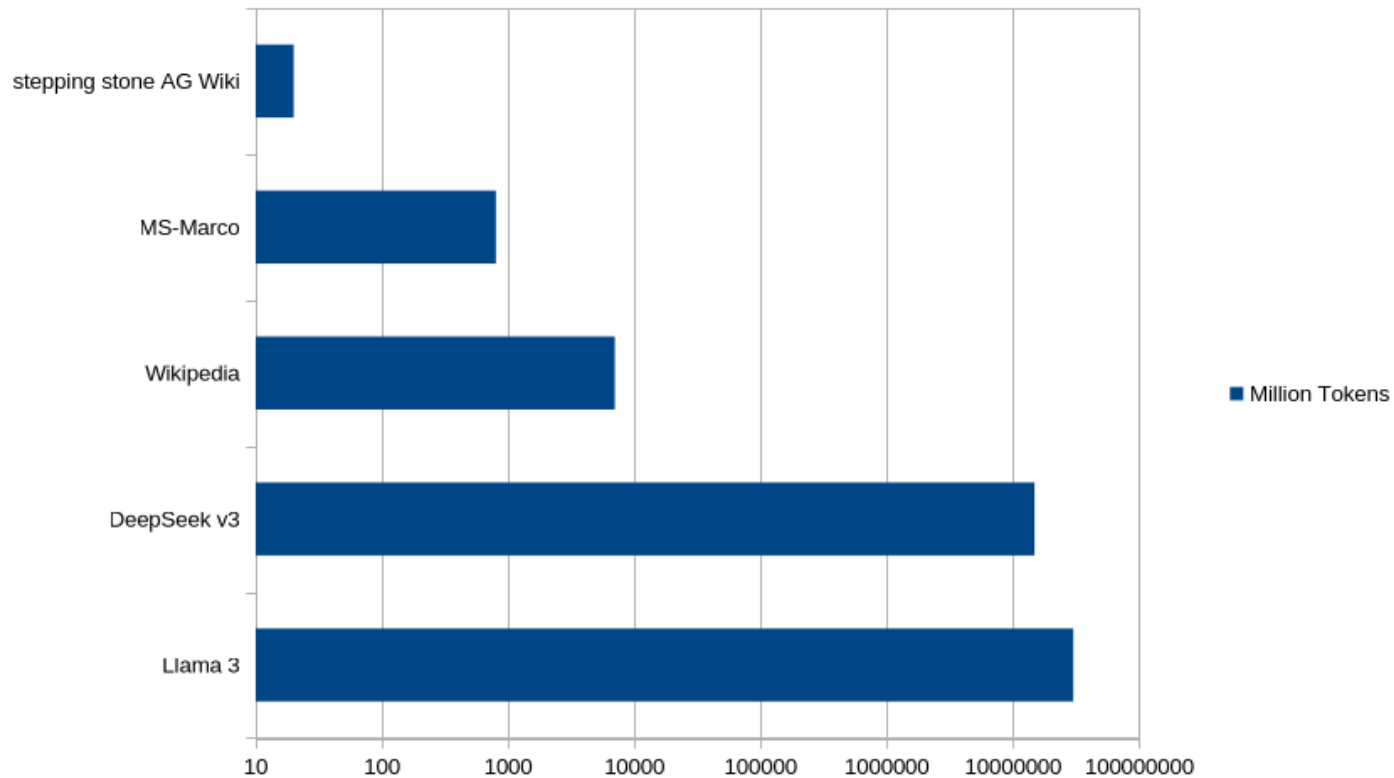
Create OpenStack VM page

Create a transcript page YYYY-MM-DD  
<Category|Company>: <Subject> .



# Extracting data

# Scale



# Extraction



- Get all the data
- Meta data is very important
- Chose an appropriate data format

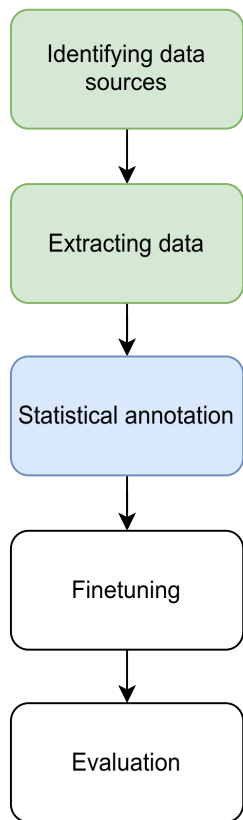
# Example: Git-MediaWiki



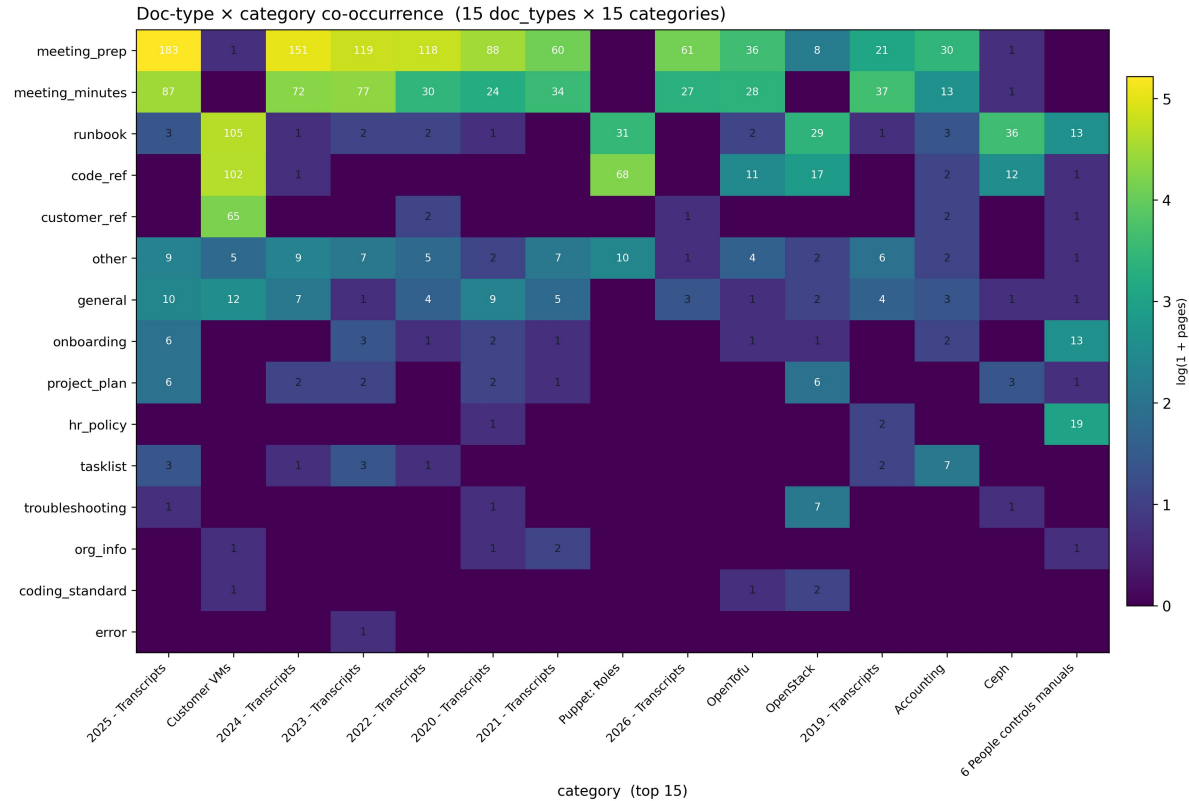
- Git-MediaWiki
  - Converts MediaWiki to Git repository



# Statistical annotation

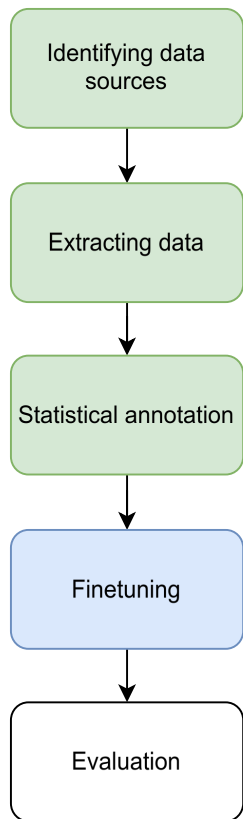


# Doc-type x Category





# Finetuning



# Finetuning: Goals



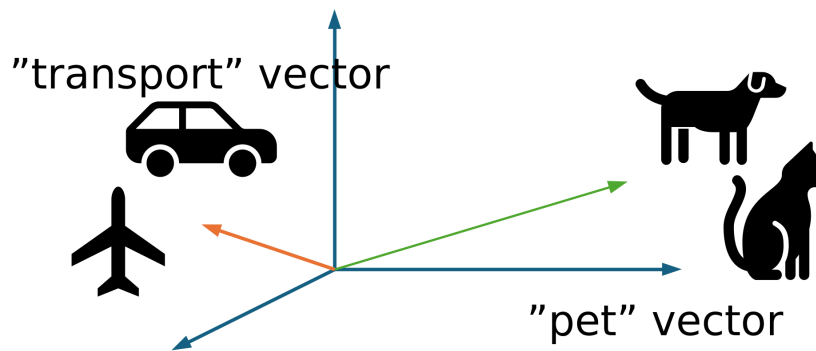
- Embedding finetuning
  - Fast search
  - Amazing retrieval
  - Use in Retrieval Augmented Generation (RAG)
- LLM Supervised Fine-Tuning (SFT)
  - Fine-Tune a Large Language Model (LLM)

# Embedding finetuning

# Embeddings in 15 seconds

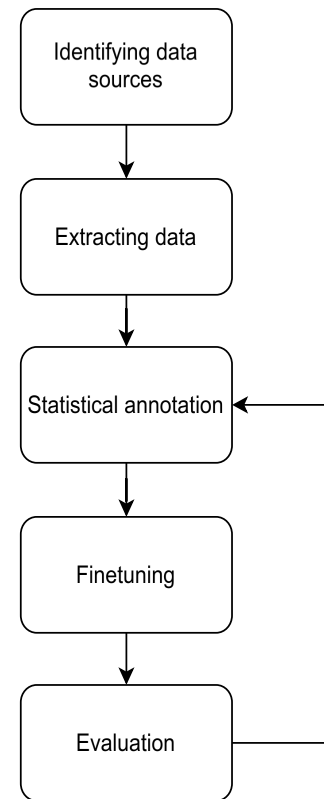


- Language lives in a high-dimensional space
- Embeddings: Map Tokens to Vectors representing relationships
- Input: Token --> Output: Closely related Tokens
  - Amazing search and retrieval

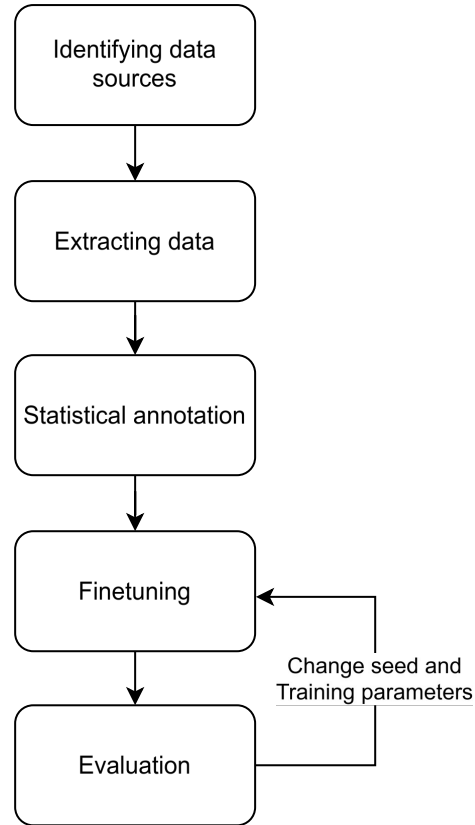


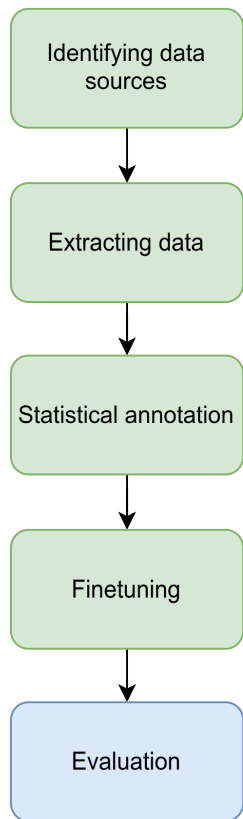
# Embedding: Process

- Use LLM to Synthesize Training data
- For each Wiki page
  - Generate Questions
- Iterate
  - Data clean up
  - Modified training behaviour



# Iterate!





# Evaluation

# Ranking

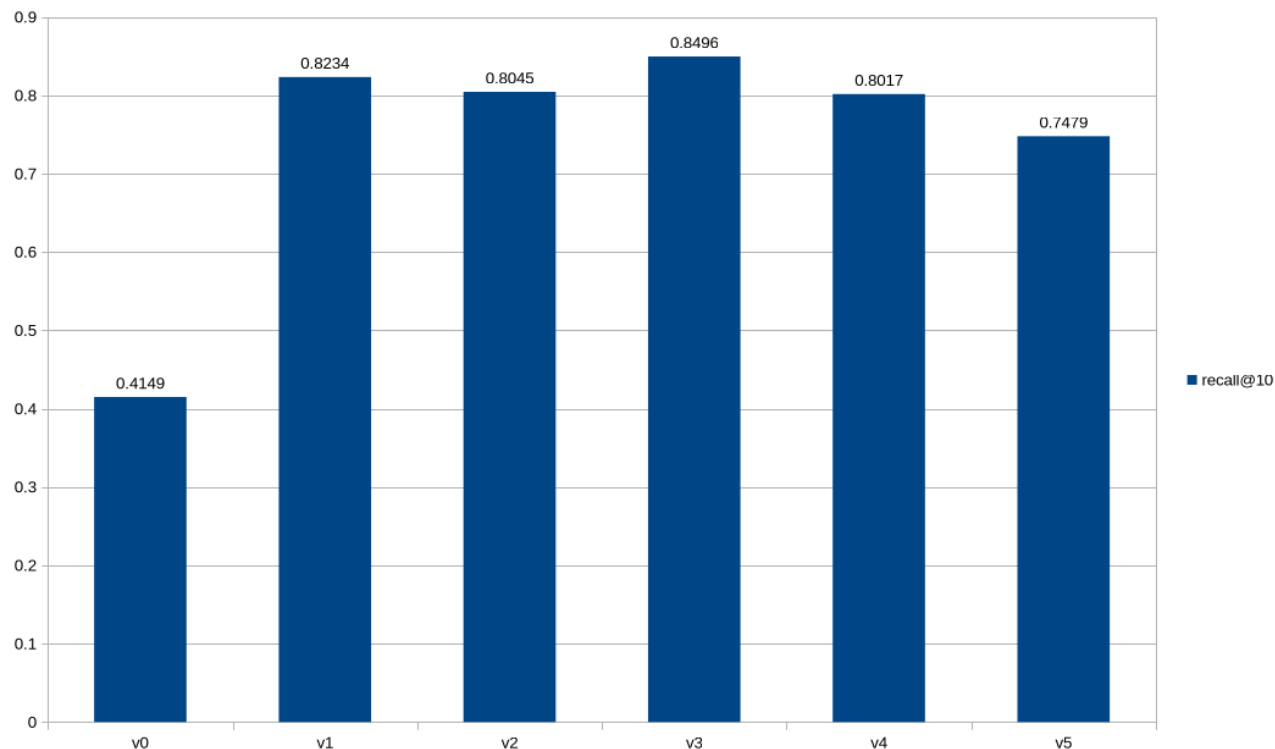


- recall @10
  - Get top-10 results from Embedding
  - Check if the right page is in there

# Embedding: Results



- e5-mli
- recall @10



# Embedding: Results



«What is the Backup policy of server sst-int-054?»

v0 Embedder:

- 1) Backup\_Policies\_Overview.mw
- 2) 2024-01-15\_Maintenance\_Log.mw
- 3) sst-int-004\_Setup\_Notes.mw
- 4) Host\_Naming\_Convention.mw

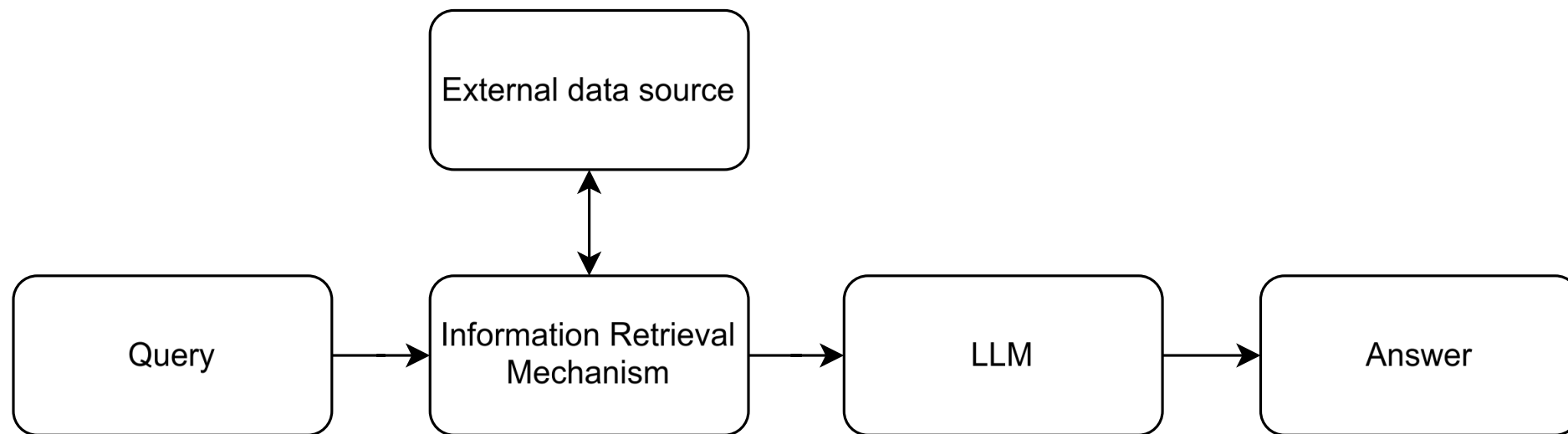
fine-tuned embedder (e5-mli, v3)

- 1) sst-int-054\_Backup\_Policy.mw
- 2) sst-int-054\_Host\_Config.mw
- 3) Backup\_Policies\_Overview.mw
- 4) sst-int-054\_Monitoring.mw

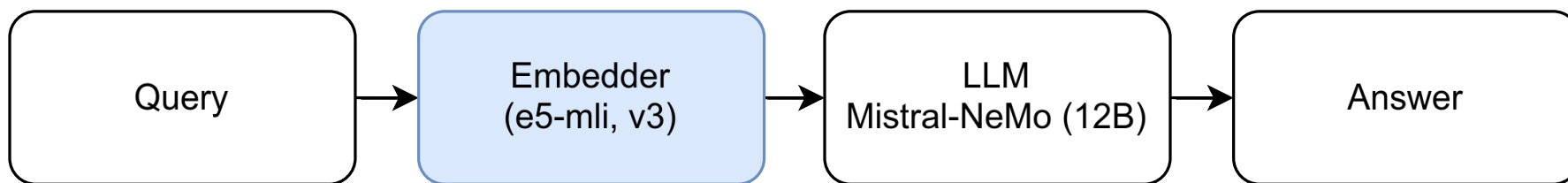
# RAG in 15 seconds



- Retrieval-Augmented Generation

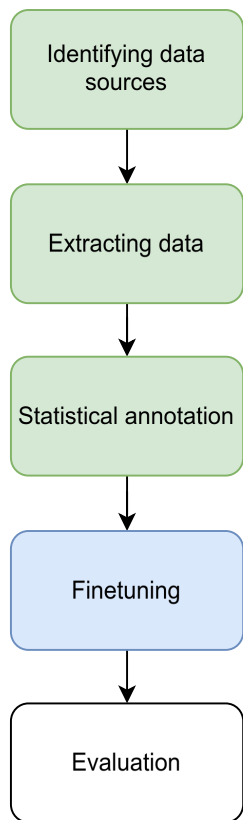


# Embedding: Use in RAG





# LLM Finetuning



# SFT: Overview

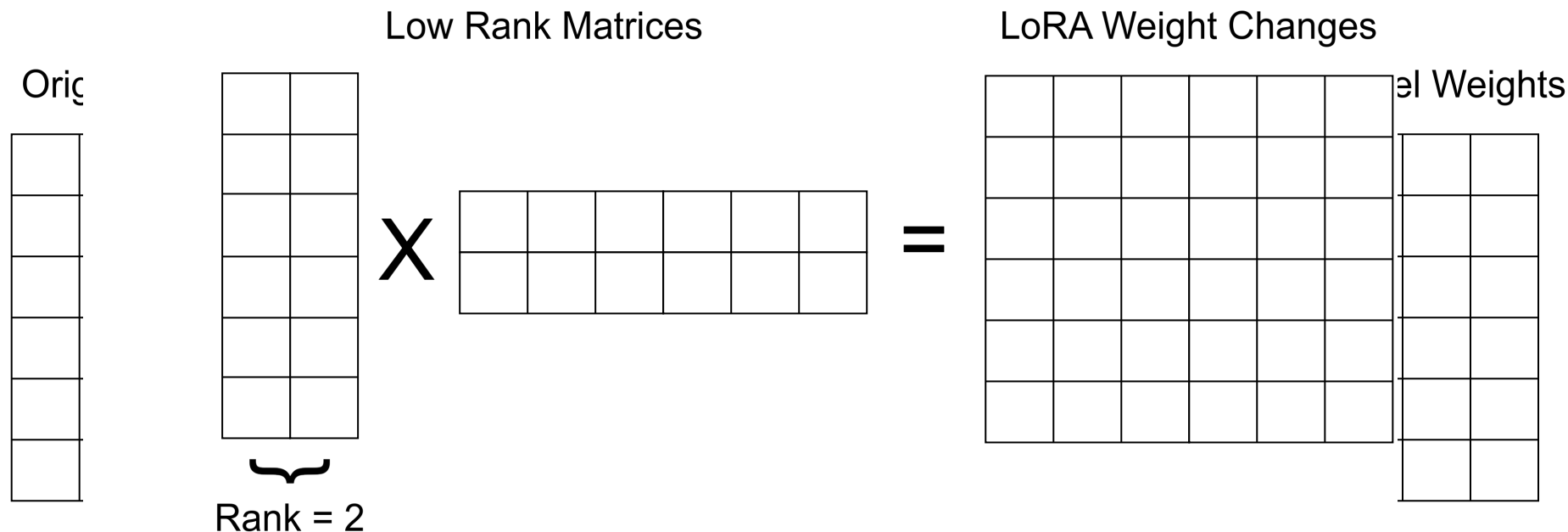


- On two different base models
  - Qwen 2.5 (3B)
  - Mistral-NeMo (12B)
- Creating an adapter
  - Using **Low-Rank Adaption** (LoRA)

# LoRA in 15 seconds



- Rank=64 --> ~0.1% (on 12B)



# Data preparation

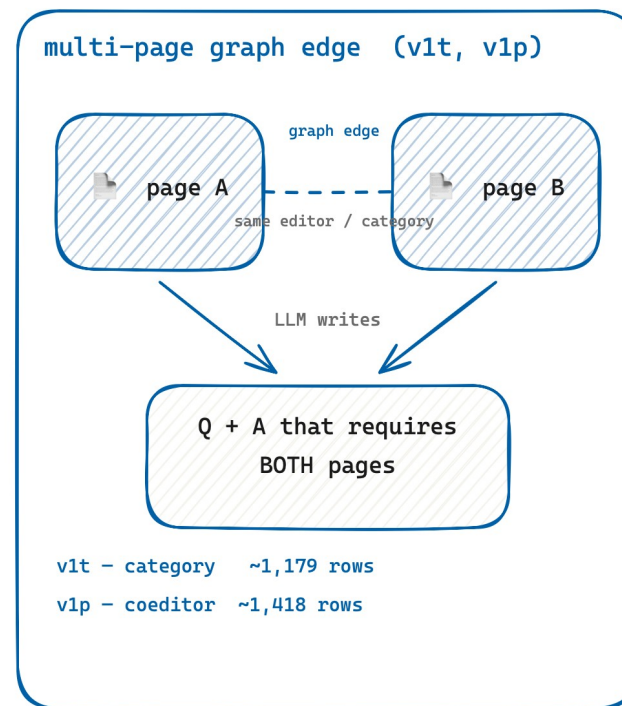
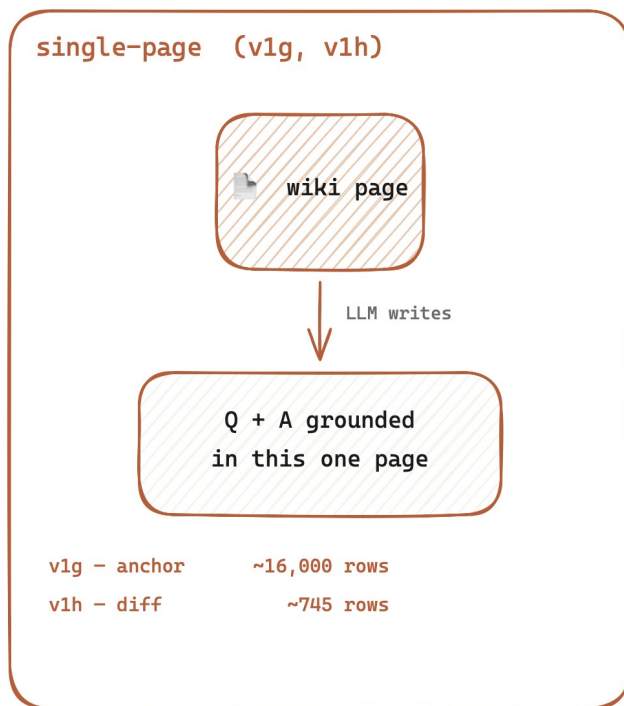


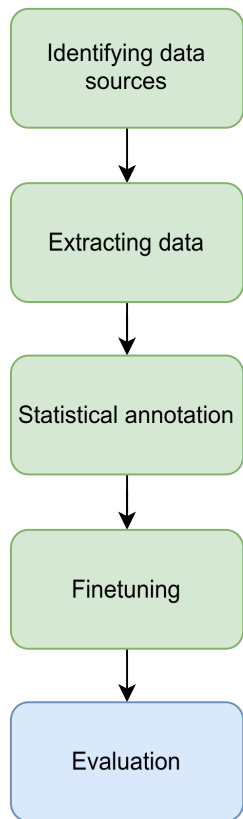
- Synthesize question / answer pairs
  - In human language
- Use LLM to do so
- Based on Wiki data
- Try different approaches to synthesis
  - We've tried and compared 16 sets of data

# SFT: Using connectivity



single-page vs multi-page graph-edge synthesis





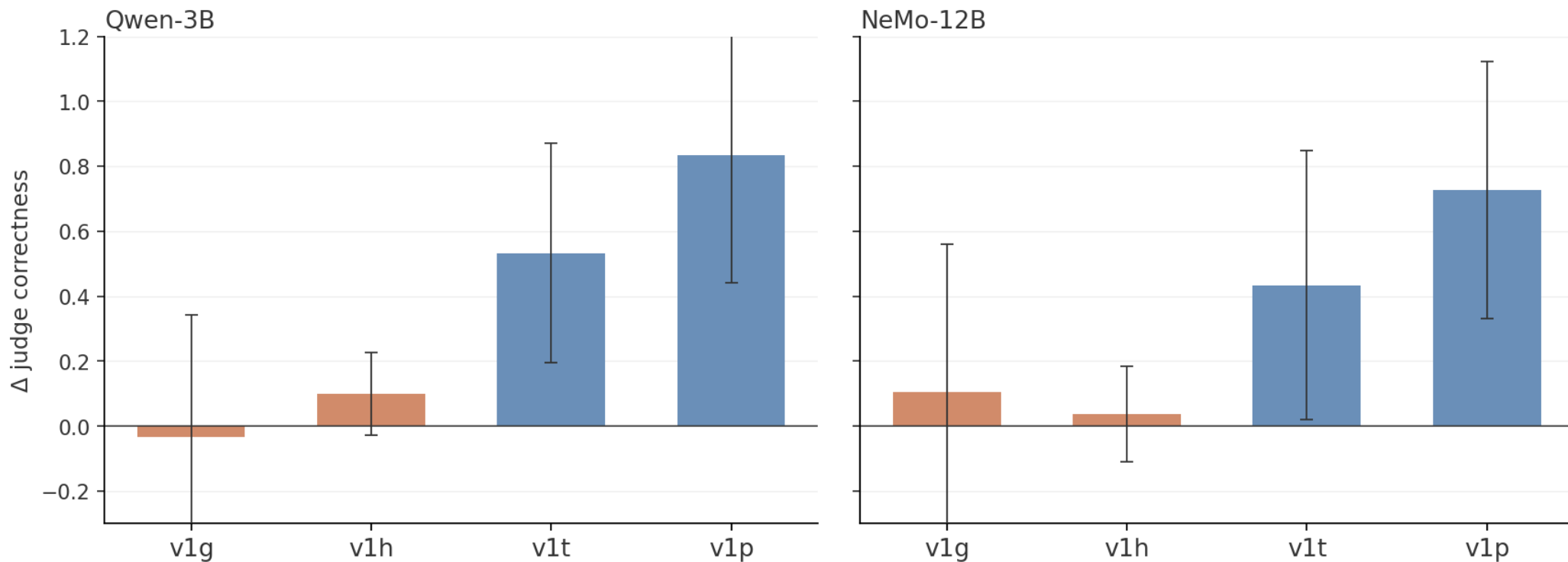
# Evaluation

# Judge

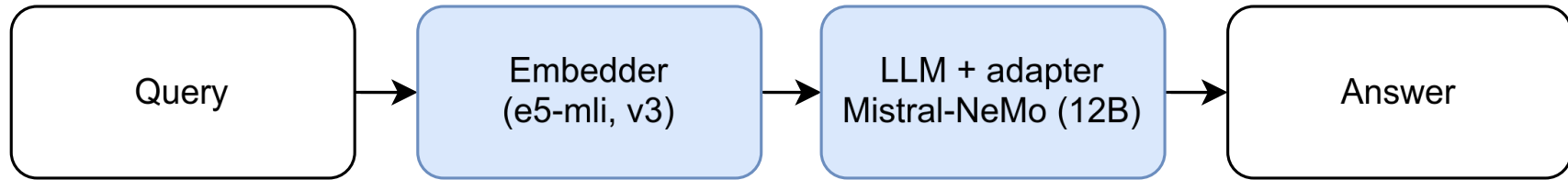


- Answers in human language
- Use LLM to rate answers

# SFT: Results



# Final pipeline



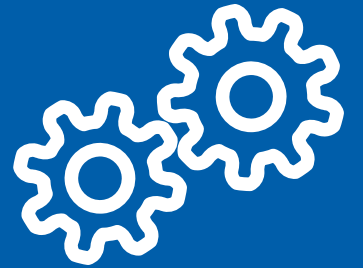
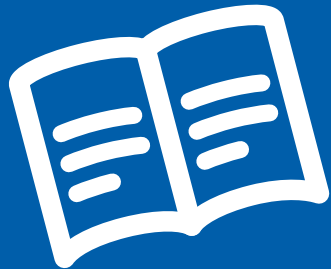
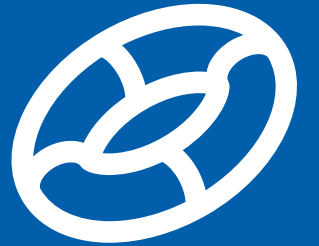
# Conclusion

# Conclusion



- Make the most of the available data
- Synthesize as much as possible
- Use LLMs for automation
- Embedding: Easy, powerful, versatile
- Try different approaches, iterate
- Not everything you read applies 1:1

Questions?



# Save the Date



What events have we planned in the field of Artificial intelligence?

- Schichtwechsel: Friday, the 4th of September 2026
- DINAcon: Wednesday, the 18th of November 2026

# Links



- Our website
  - <https://www.stepping-stone.ch/>
- Upcoming events
  - <https://www.stepping-stone.ch/schichtwechsel/>
  - <https://dinacon.ch/>



# **stepping stone AG**

Wasserwerksgasse 7  
3011 Bern

Telefon: +41 31 332 53 63  
[www.stepping-stone.ch](http://www.stepping-stone.ch)  
[info@stepping-stone.ch](mailto:info@stepping-stone.ch)