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

Niklaus Hofer

Uphill Conf

8th 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

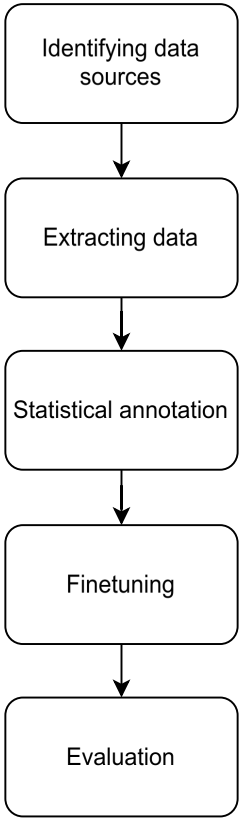
- Niklaus Hofer
- Joined in 2016
- CTO

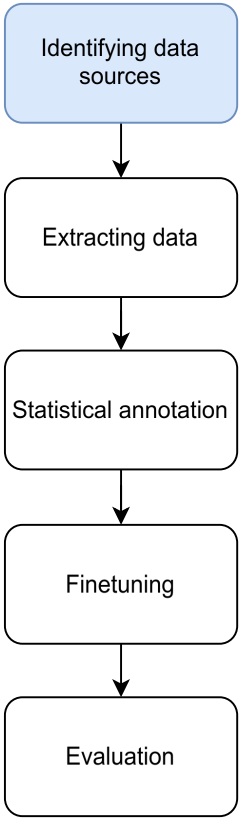
Introduction

Serving data to AI

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

Pipeline


Identifying data sources
Extracting data
Statistical annotation
Finetuning
Evaluation

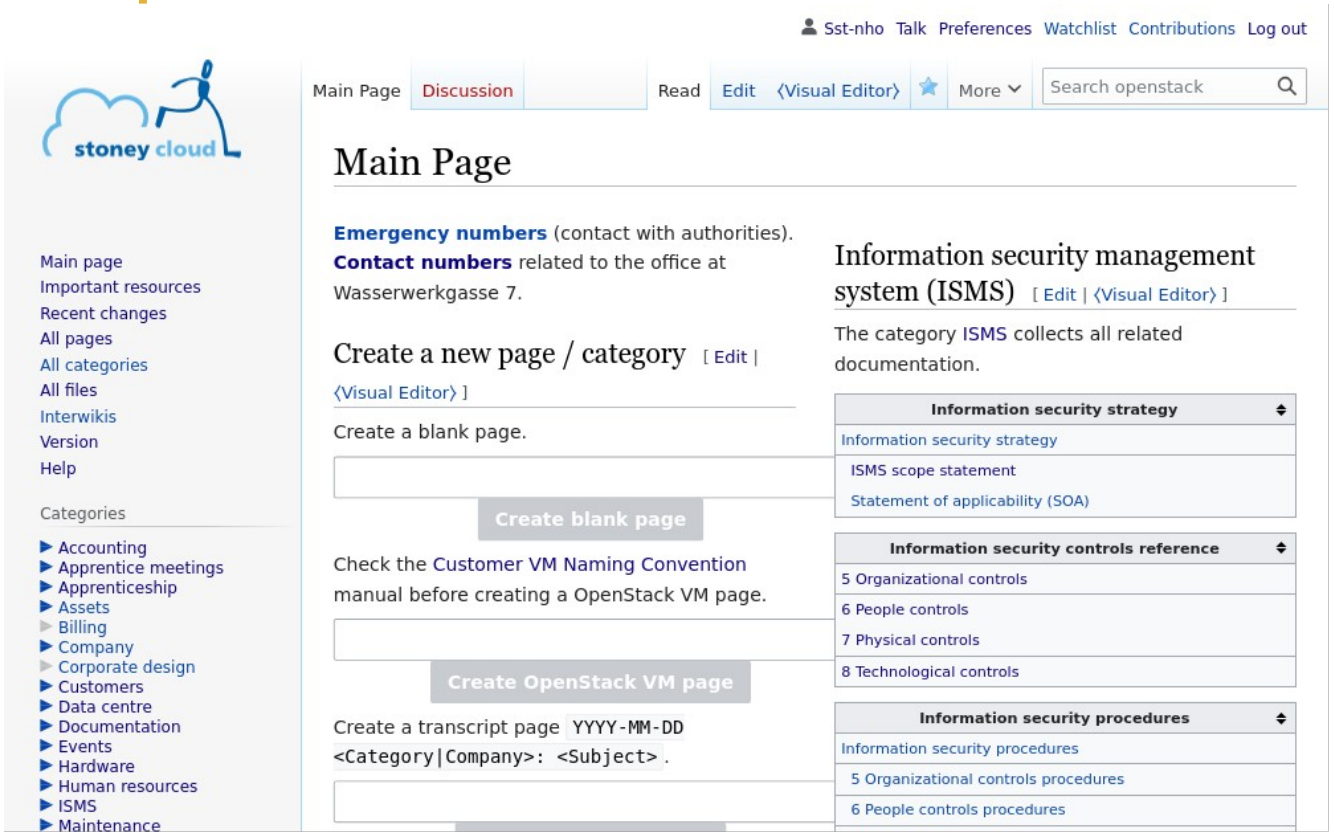


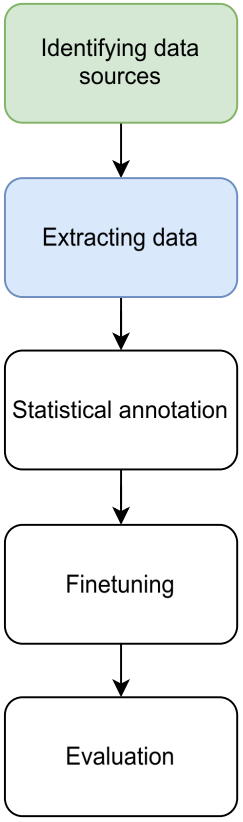
Identifying data sources

Data sources

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

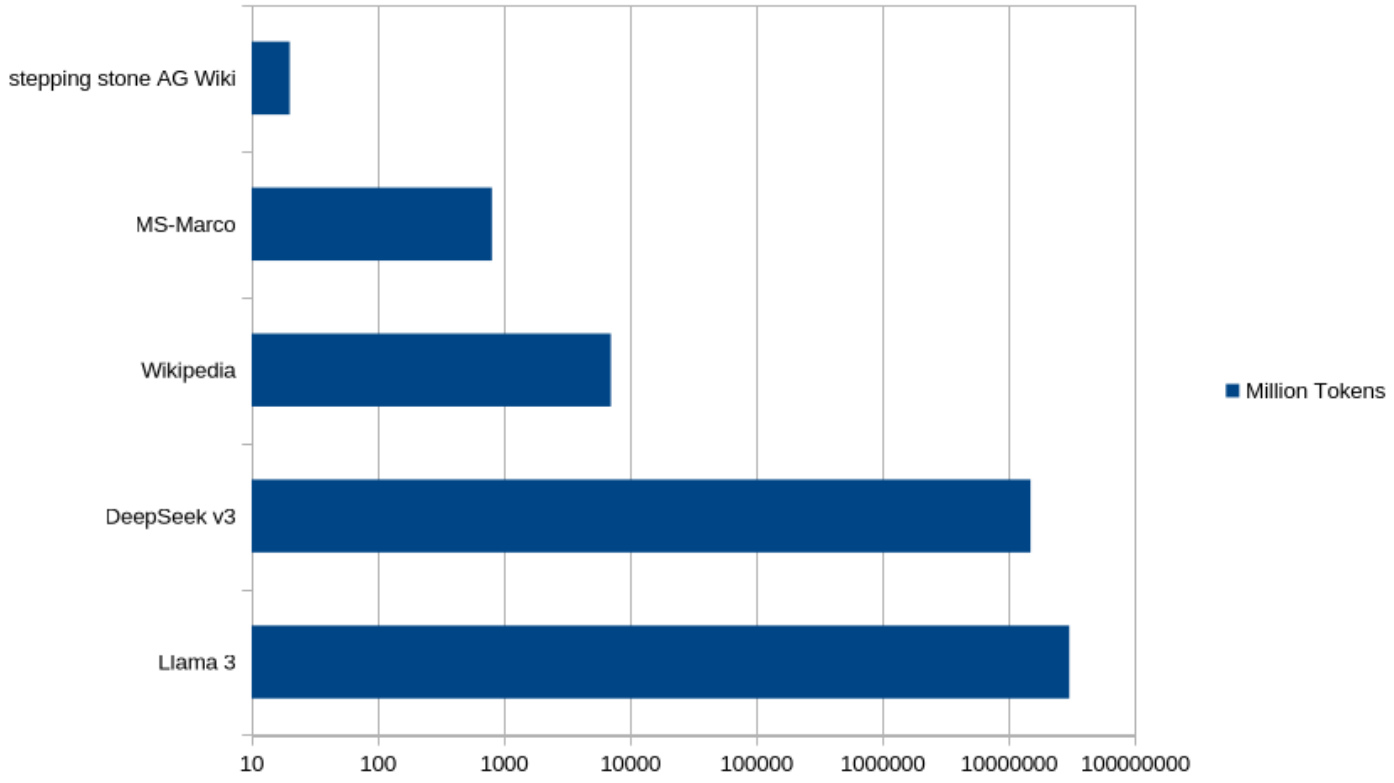
Example: MediaWiki





Extracting data

Scale

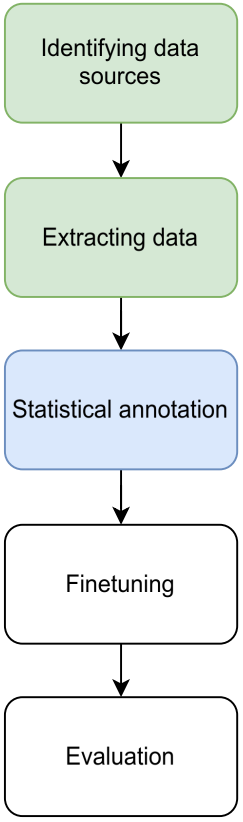

stepping stone AG Wiki
MS-Marco
Wikipedia
DeepSeek v3
Llama 3
10
100
1000
10000
100000
1000000
10000000
100000000
Million Tokens

Extraction

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

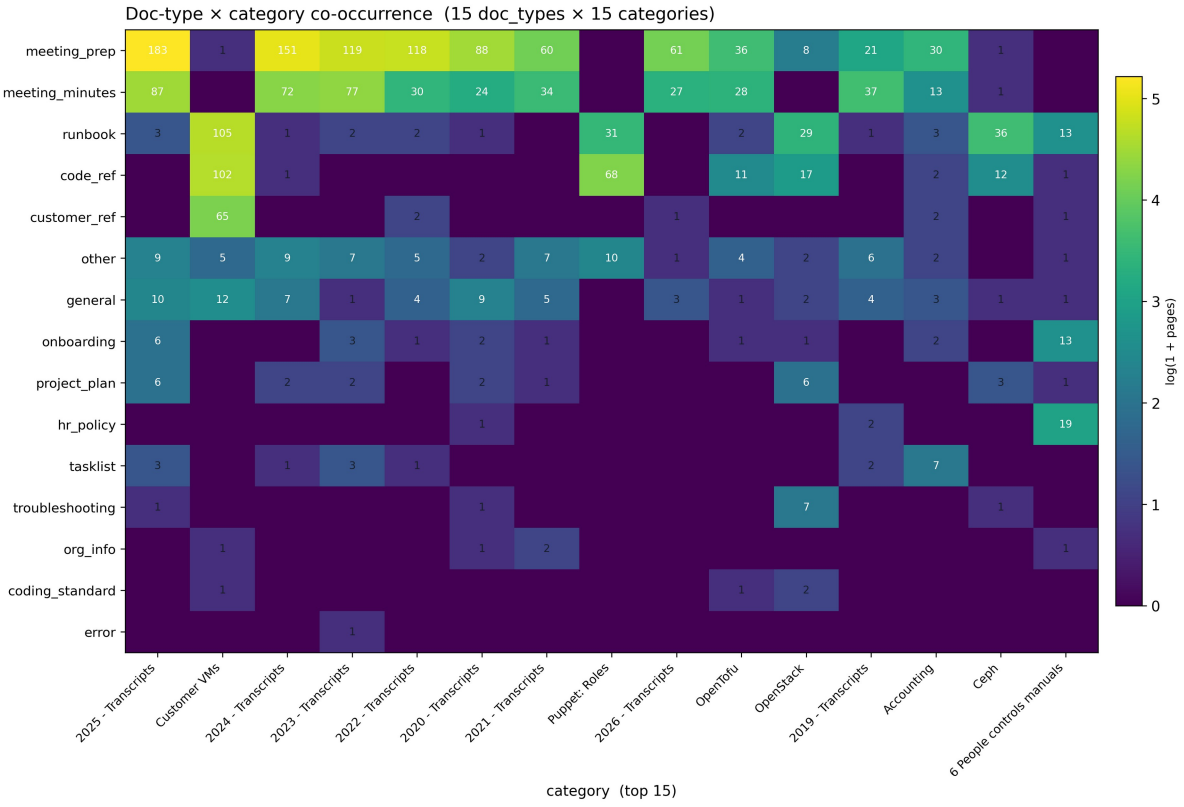
Example: Git-MediaWiki

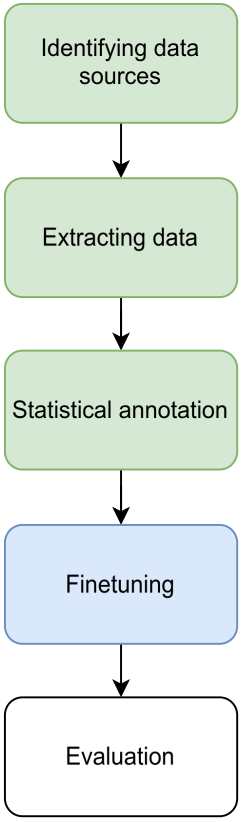
- Git-MediaWiki
  - Converts MediaWiki to Git repository



Statistical annotation

Doc-type x Category




Identifying data sources
Extracting data
Statistical annotation
Finetuning
Evaluation

Finetuning

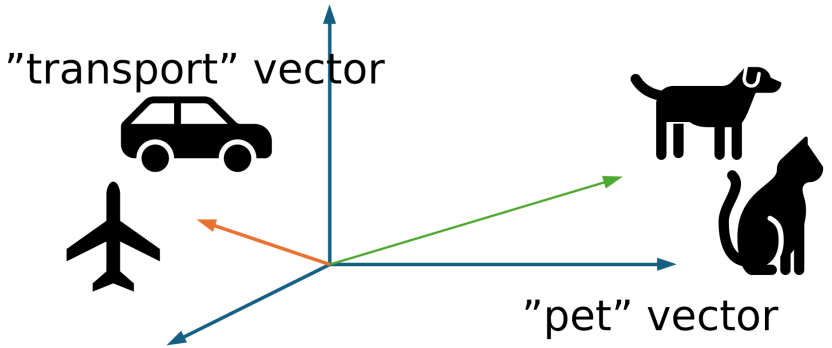
Finetuning: Goals

- 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

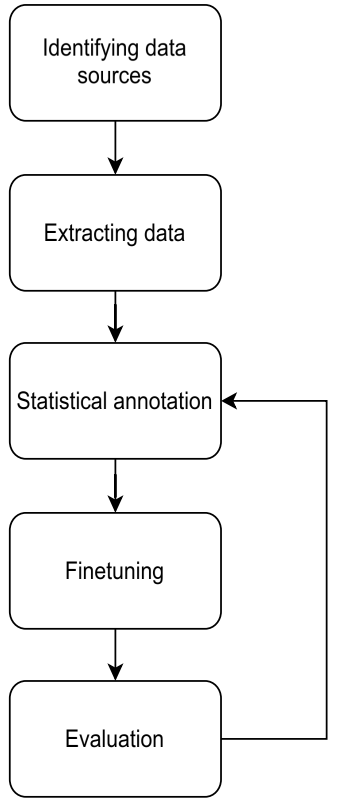
Embeddings in 15 seconds

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

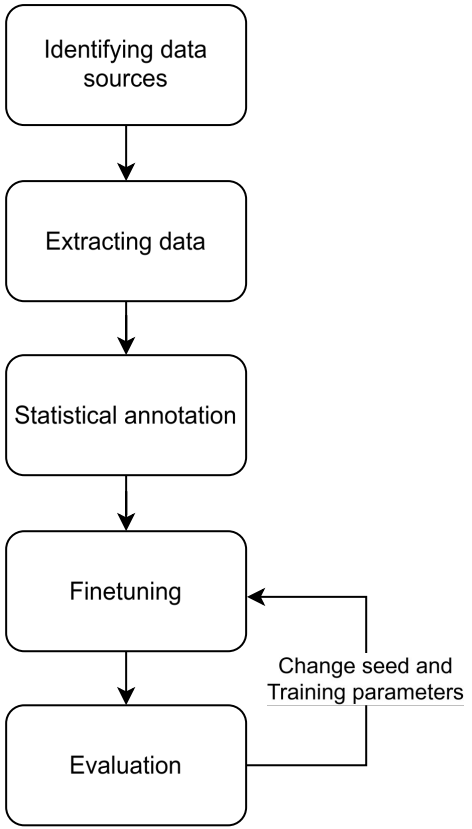


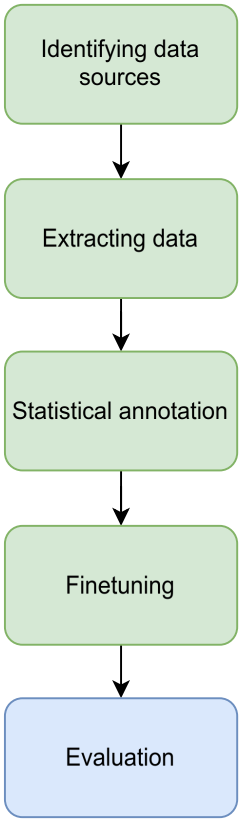
Embedding: Process

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



Iterate!


Identifying data sources
Extracting data
Statistical annotation
Finetuning
Change seed and Training parameters
Evaluation


Identifying data sources
Extracting data
Statistical annotation
Finetuning
Evaluation

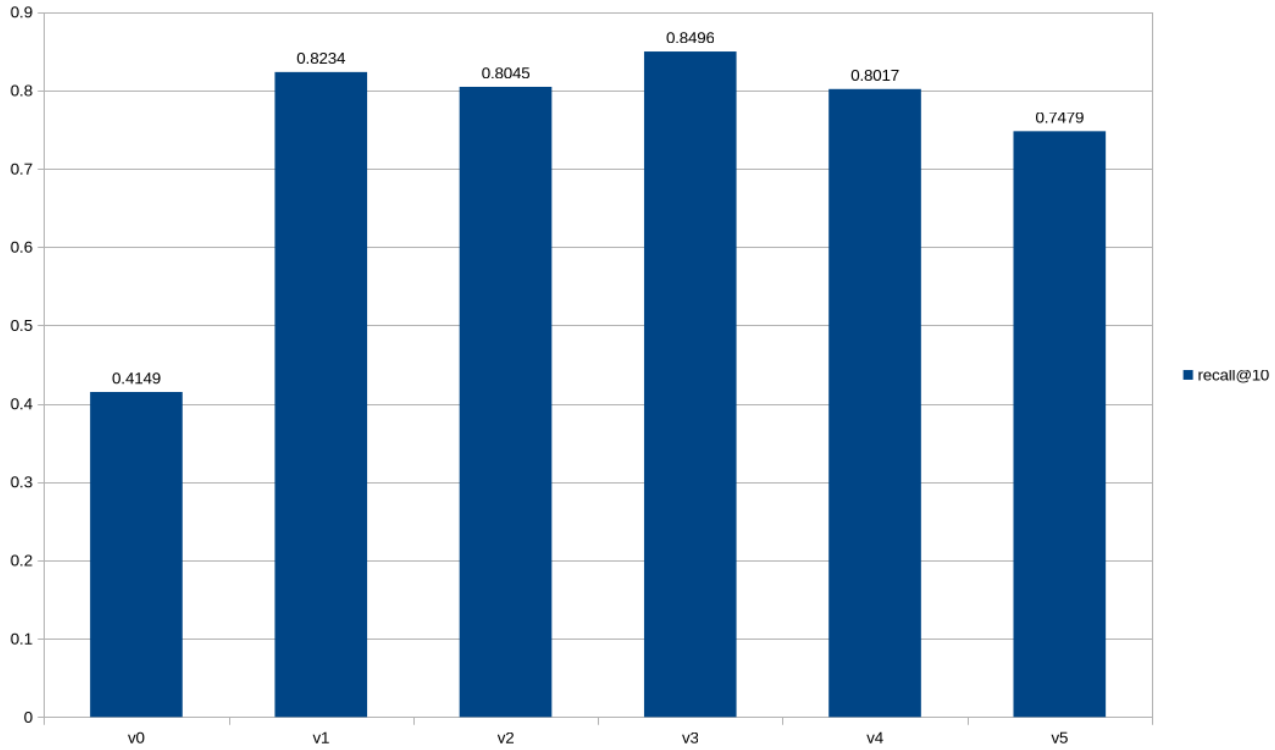
Evaluation

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

stepping stone

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

v0 Embedder:

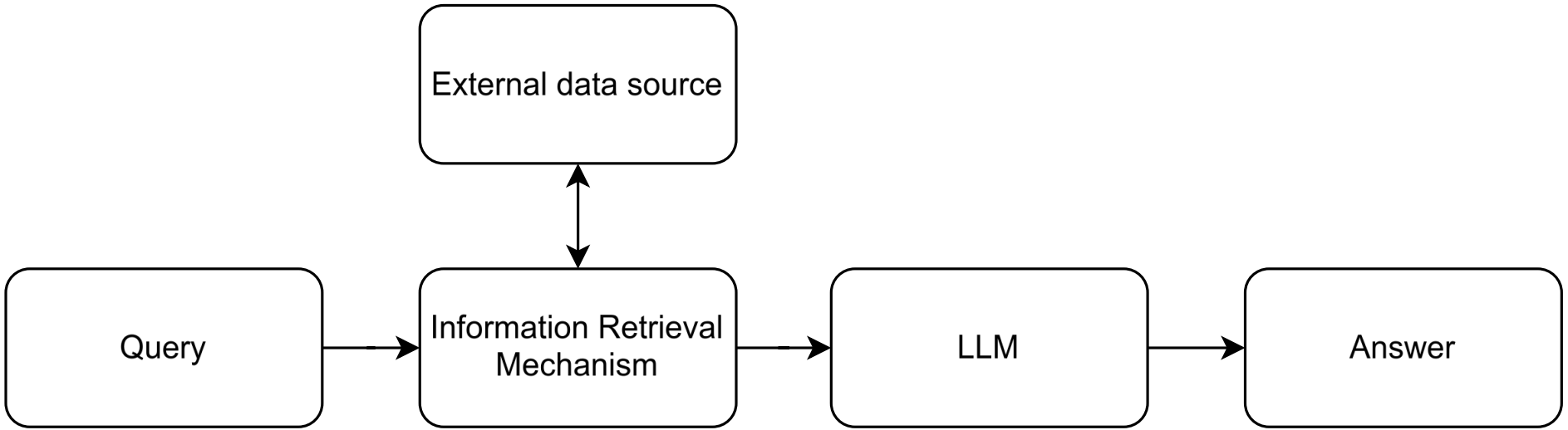
1) Backup_Policies_Overview.mw
2) 2024-01-15_Maintenance_Log.mw
3) sst-int-004_Setup_Notes.mw
4) Host_Naming_Convention.mw

fine-tuned embedder (e5-mli, v3)

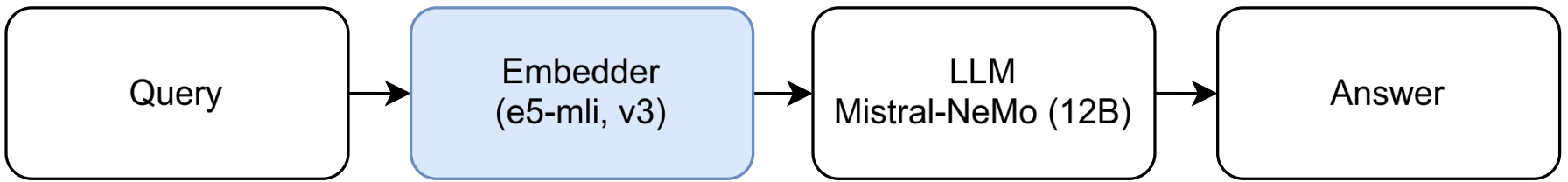
1) sst-int-054_Backup_Policy.mw
2) sst-int-054_Host_Config.mw
3) Backup_Policies_Overview.mw
4) sst-int-054_Monitoring.mw

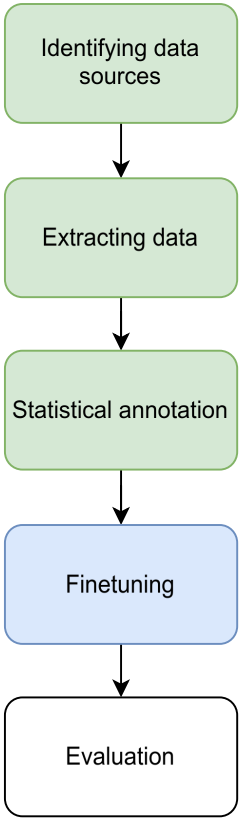
RAG in 15 seconds

- Retrieval-Augmented Generation



Embedding: Use in RAG





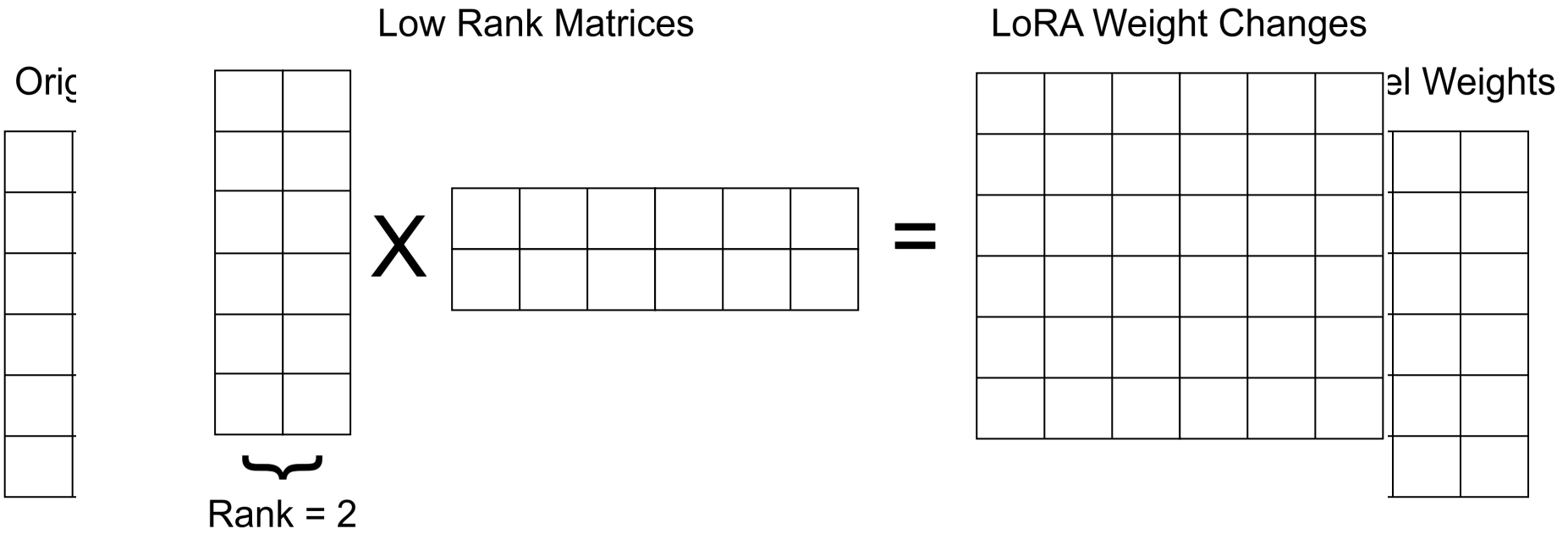
LLM Finetuning

SFT: Overview

- On two different base models
  - Qwen 2.5 (3B)
  - Mistral-NeMo (12B)
- Creating an adapter
  - Using **Lo**w-**R**ank **A**daption (LoRA)

LoRA in 15 seconds

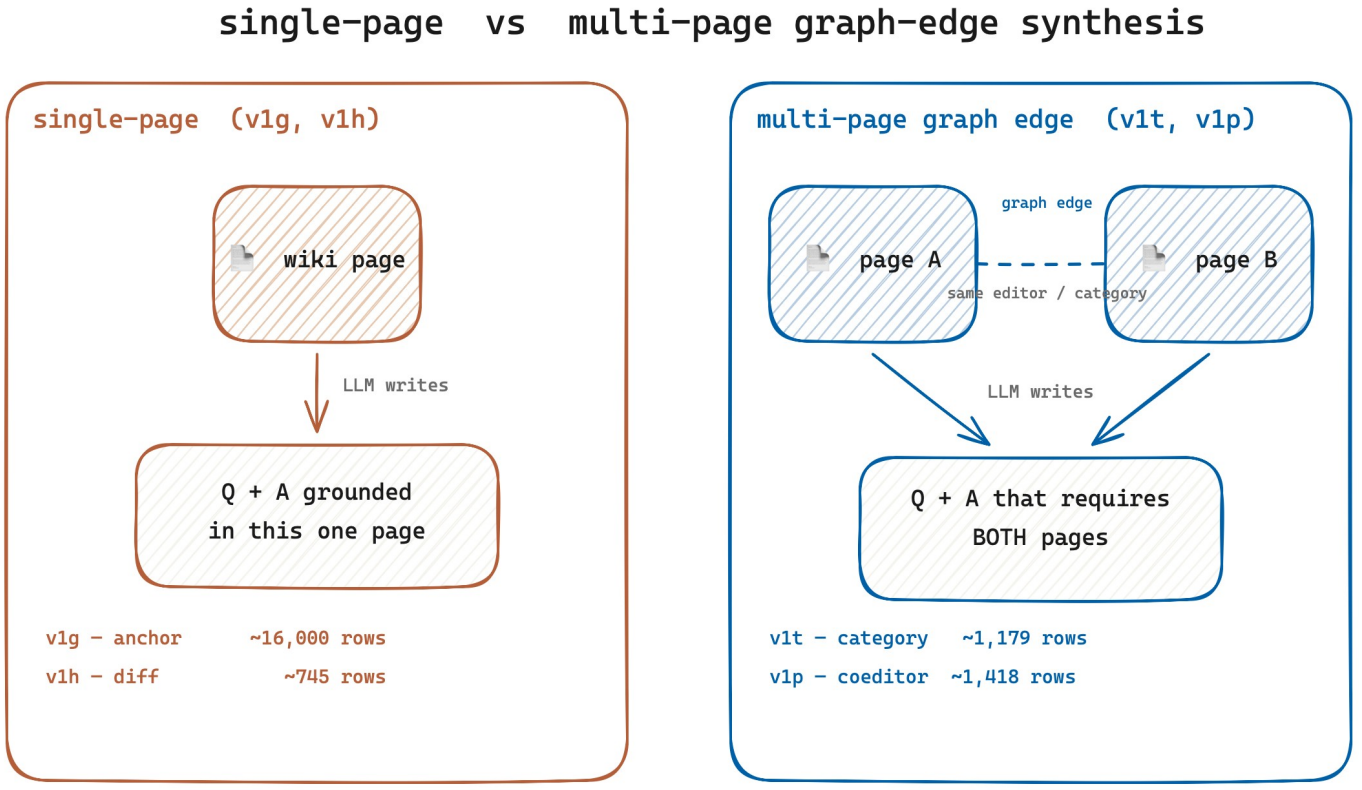
- Rank=64 --> ~0.1% (on 12B)

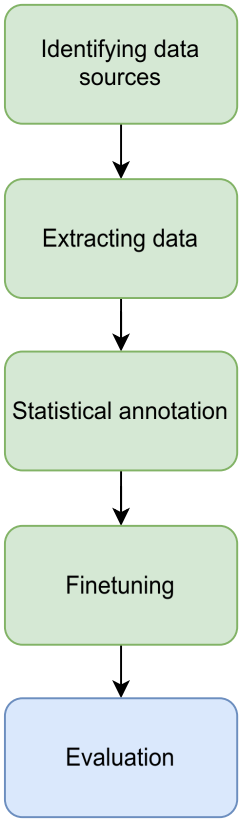


Data preparation

- 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



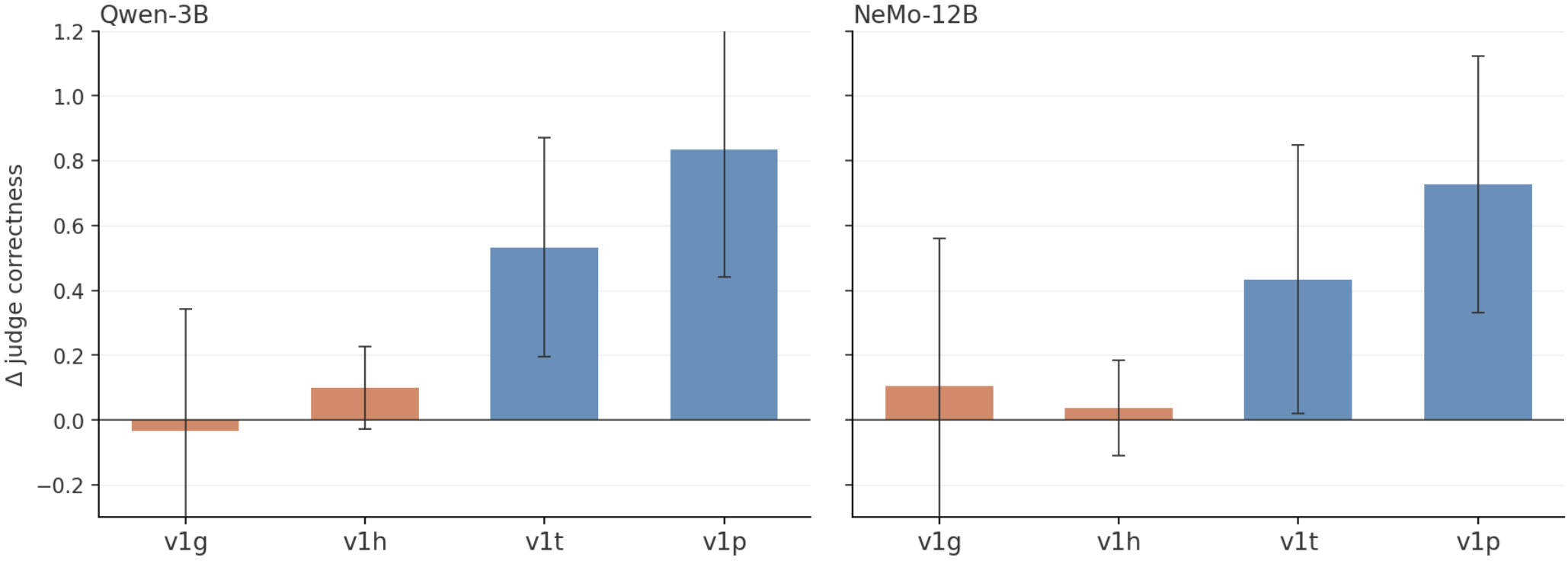

Identifying data sources
Extracting data
Statistical annotation
Finetuning
Evaluation

Evaluation

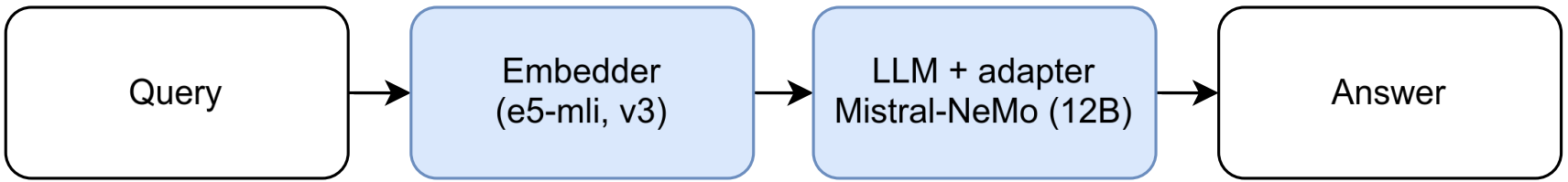
Judge

- Answers in human language
- Use LLM to rate answers

SFT: Results


Qwen-3B
NeMo-12B
Δ judge correctness
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
v1g
v1h
v1t
v1p

Final pipeline



Conclusion

Conclusion

stepping stone

- 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
Wasserwerkgasse 7
3011 Bern

Telefon: +41 31 332 53 63
www.stepping-stone.ch
info@stepping-stone.ch