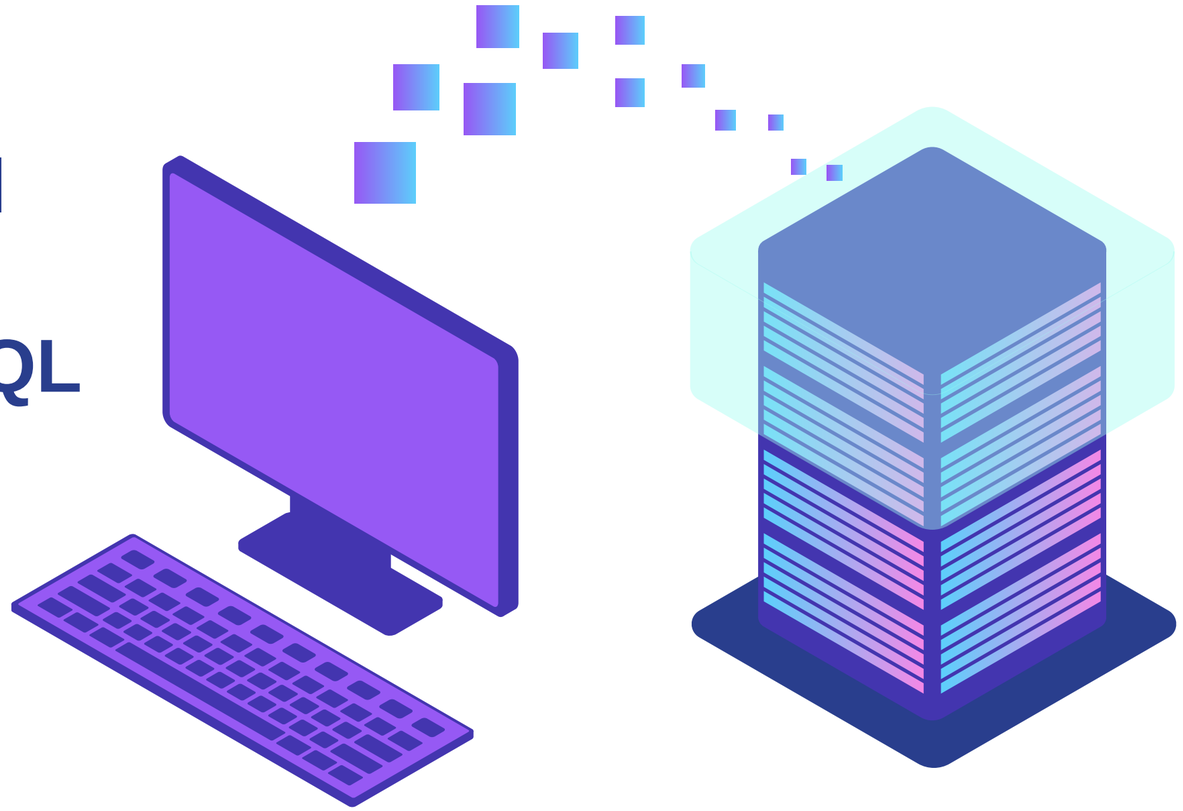
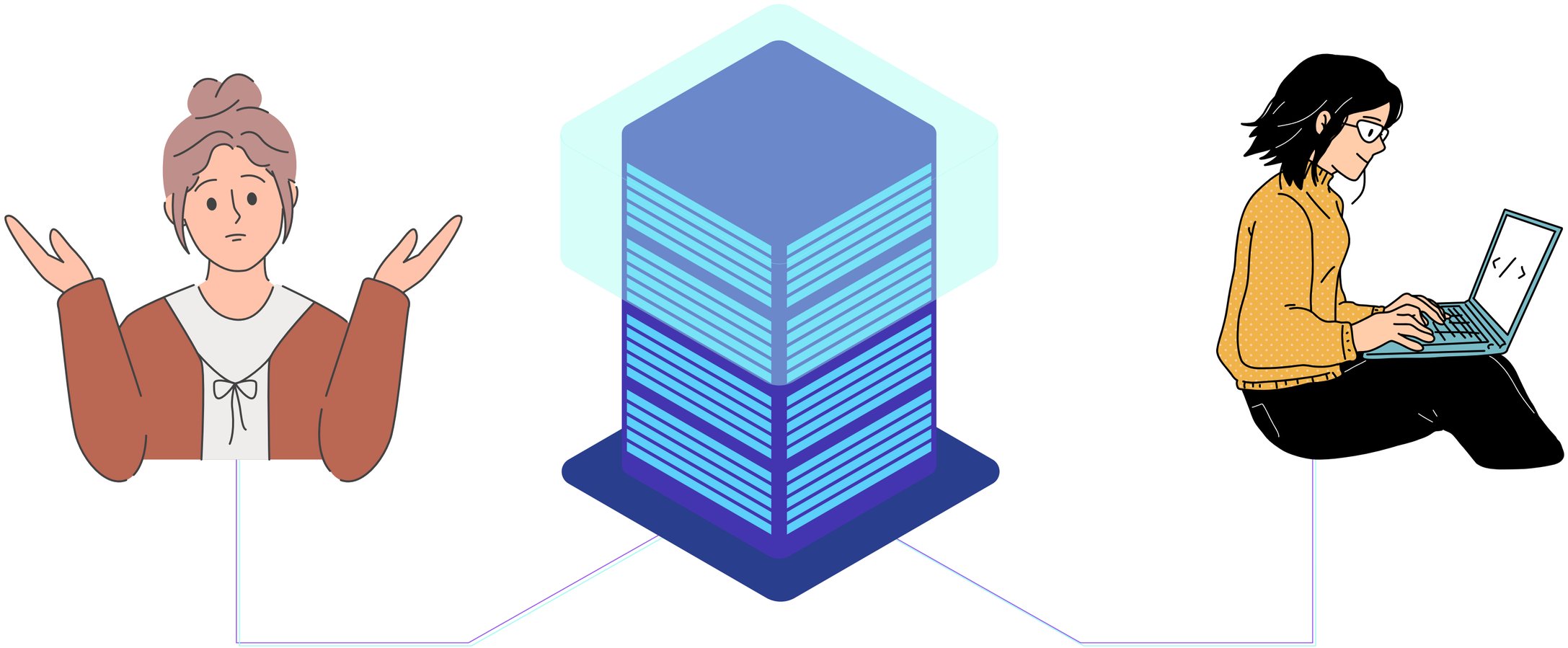
Fine-Tuning a T5-Small Model for Multitask Text-to-SQL Generation within the MAC-SQL Paradigm

By Elsa Ruiz



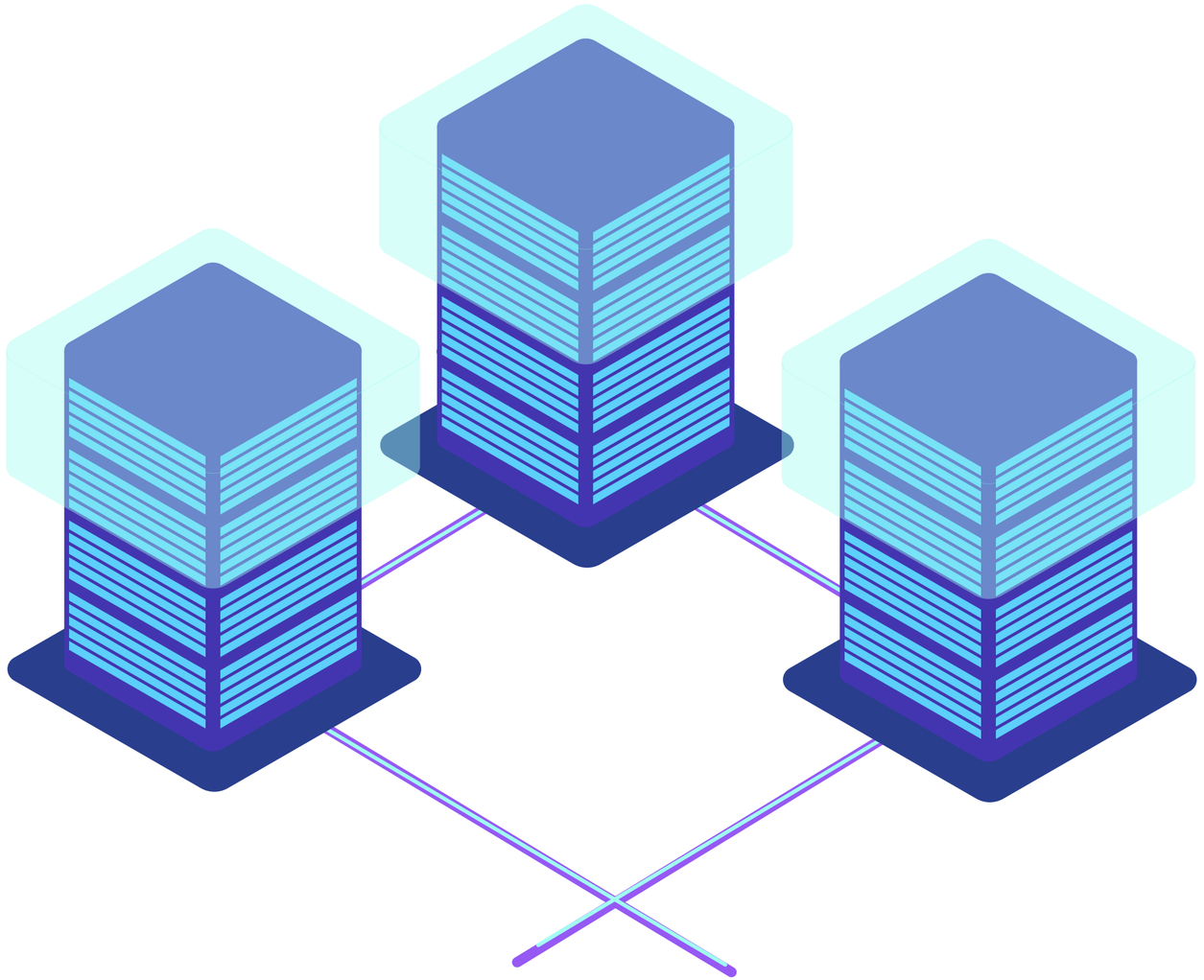
Relevance



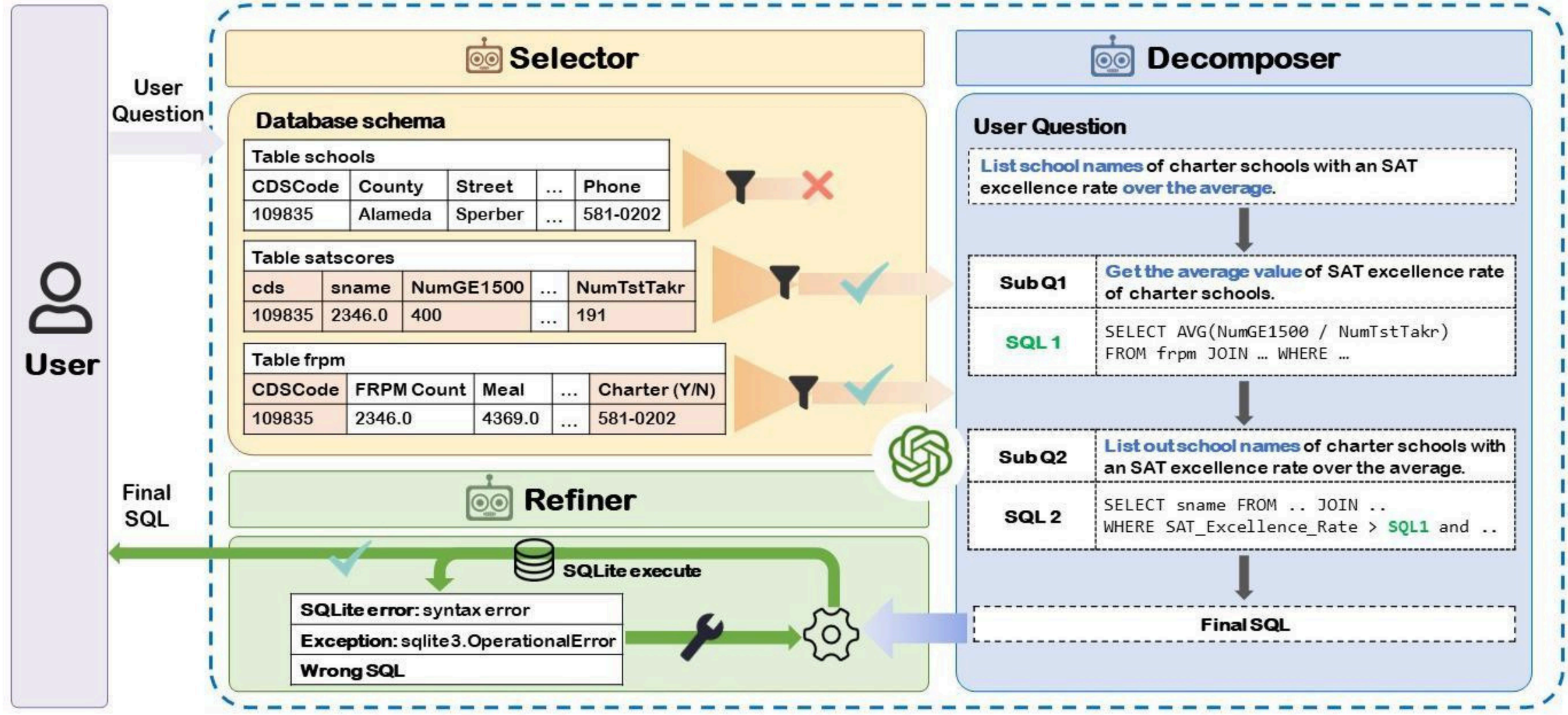
Multi Agent Collaborative SQL

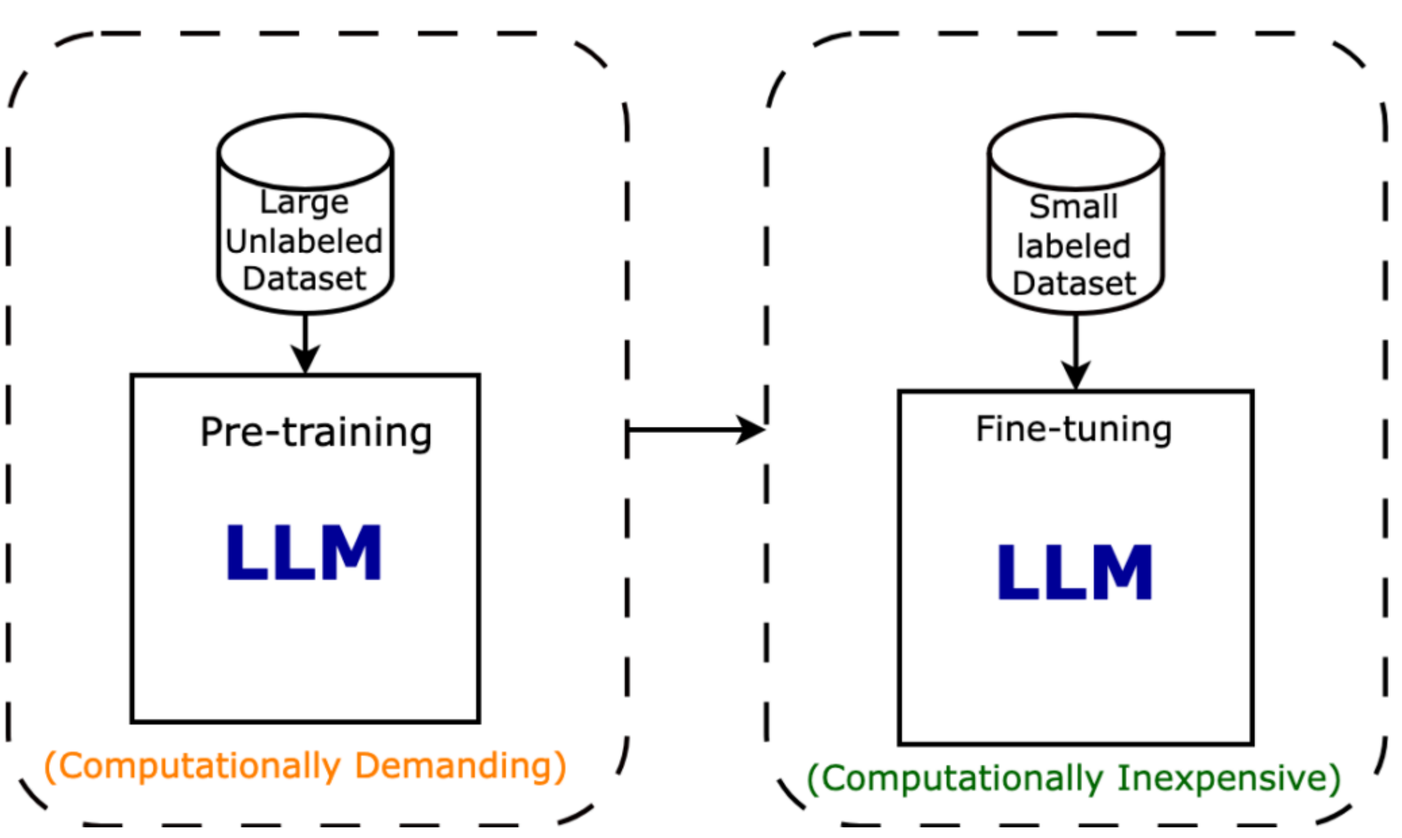
2. Decomposer

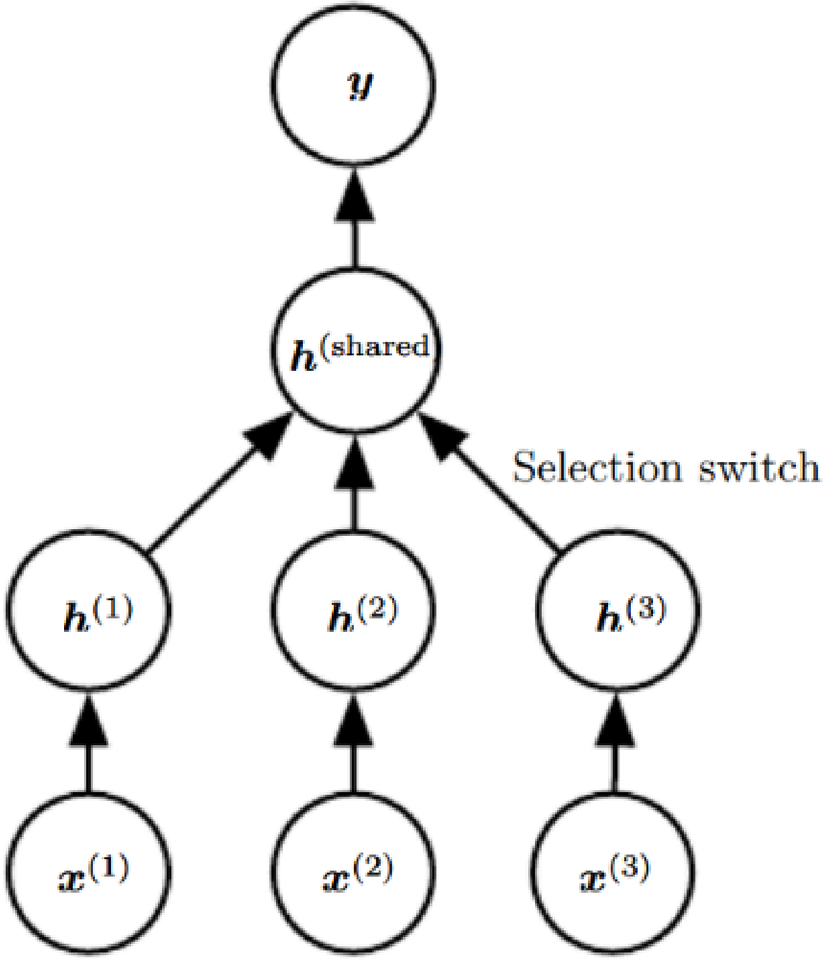
1. Selector



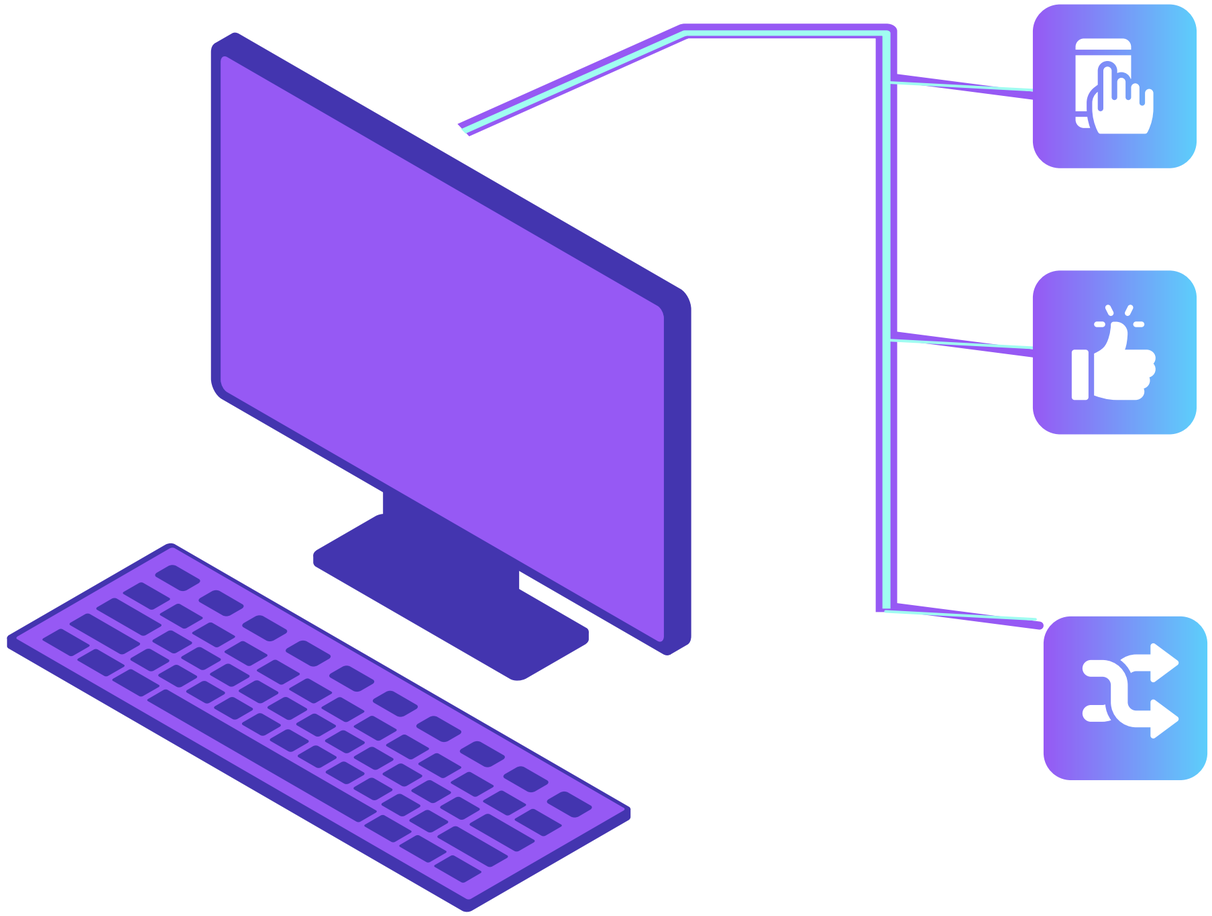
3. Refiner


User
User Question
Selector
Database schema
Table schools
CDSCode | County | Street | ... | Phone
109835 | Alameda | Sperber | ... | 581-0202
Table satscores
cds | sname | NumGE1500 | ... | NumTstTakr
109835 | 2346.0 | 400 | ... | 191
Table frpm
CDSCode | FRPM Count | Meal | ... | Charter (Y/N)
109835 | 2346.0 | 4369.0 | ... | 581-0202
Decomposer
User Question
List school names of charter schools with an SAT excellence rate over the average.
Sub Q1
Get the average value of SAT excellence rate of charter schools.
SQL 1
SELECT AVG(NumGE1500 / NumTstTakr)
FROM frpm JOIN … WHERE …
Sub Q2
List out school names of charter schools with an SAT excellence rate over the average.
SQL 2
SELECT sname FROM .. JOIN ..
WHERE SAT_Excellence_Rate > SQL1 and ..
Final SQL
Refiner
SQLite execute
SQLite error: syntax error
Exception: sqlite3.OperationalError
Wrong SQL
Final SQL


Large
Unlabeled
Dataset
Pre-training
LLM
(Computationally Demanding)
Small
labeled
Dataset
Fine-tuning
LLM
(Computationally Inexpensive)


y
$h^{(\text{shared}}$
Selection switch
$h^{(1)}$
$h^{(2)}$
$h^{(3)}$
$x^{(1)}$
$x^{(2)}$
$x^{(3)}$

Project



Goal

Enable natural language interfaces for SQL databases using a lightweight model and few hundred data to explore power of finetuning

Challenge

- Translate ambiguous natural queries into valid, executable SQL.
- Reach a good accuracy in resource limited environment

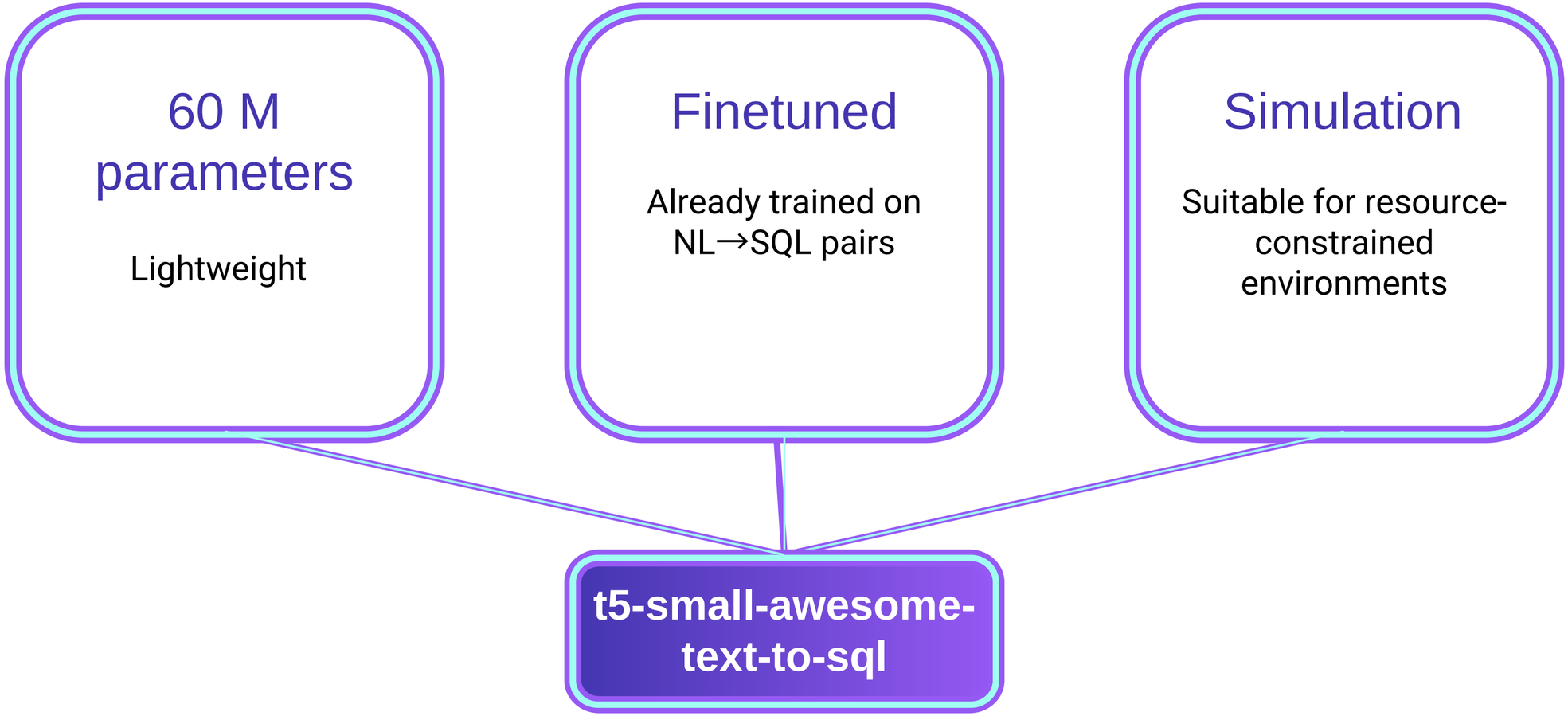
Approach

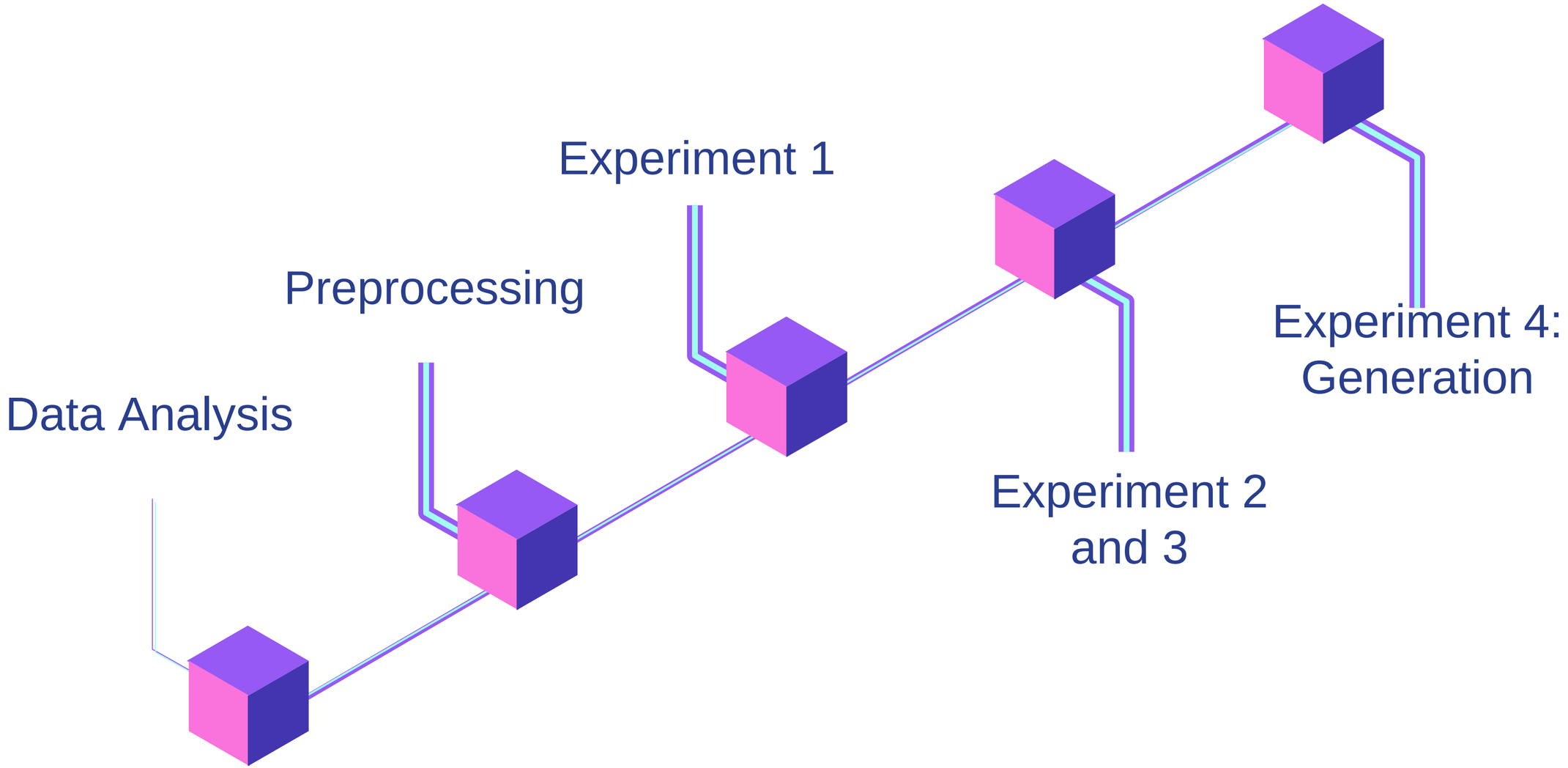
Fine-tune t5-small-awesome-text-to-sql using a multitask dataset (Selector, Decomposer, Refiner) under the MAC-SQL architecture.

Data migration process infographics

Source	Target	...

GPT-4	Propietary, Innaccessible	
SQL-Llama 7B	Code-focused, resources	
T5-Small	For experimentation faster and efficient	


60 M parameters
Lightweight
Finetuned
Already trained on NL→SQL pairs
Simulation
Suitable for resource-constrained environments
t5-small-awesome-text-to-sql


Experiment 1
Preprocessing
Data Analysis
Experiment 2
and 3
Experiment 4:
Generation

- Preprocessing
- Trainer API Hugging Face
- Initial parameters:
  - Batch size: 1
  - Epochs: 10
  - Metric for best model: ‘eval_loss’
  - Learning rate: 5e-5
  - Models and logs storage
  - bf16

Metrics

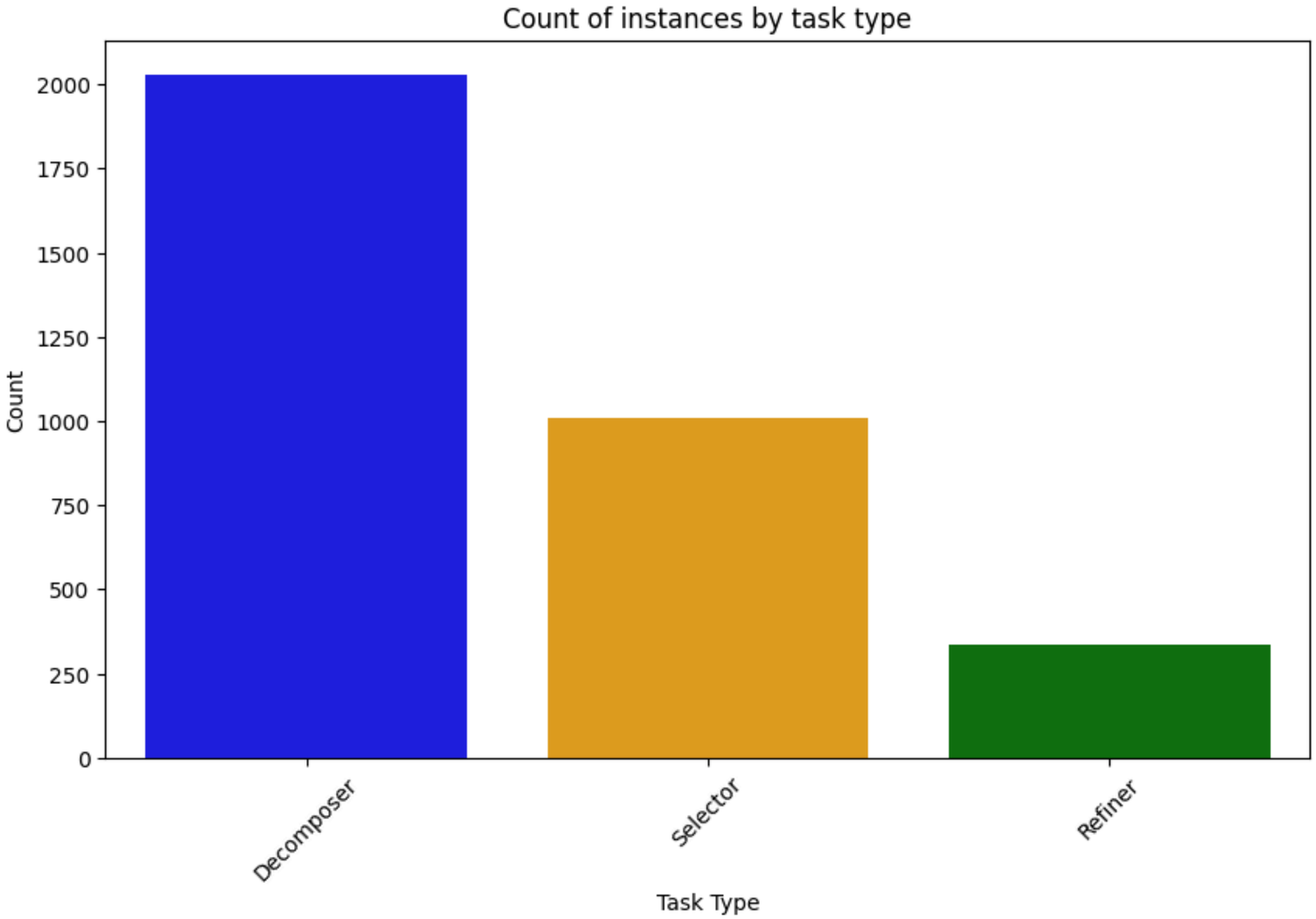
Loss: **Cross-Entropy**

Evaluation:

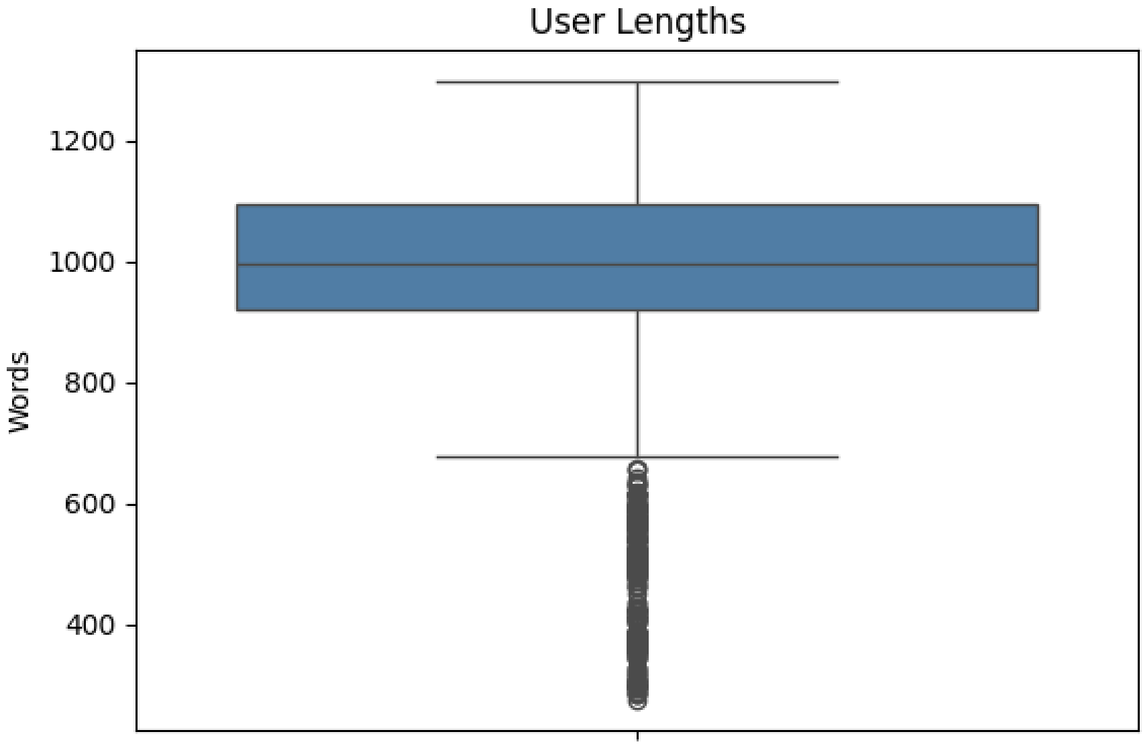
Exact Match

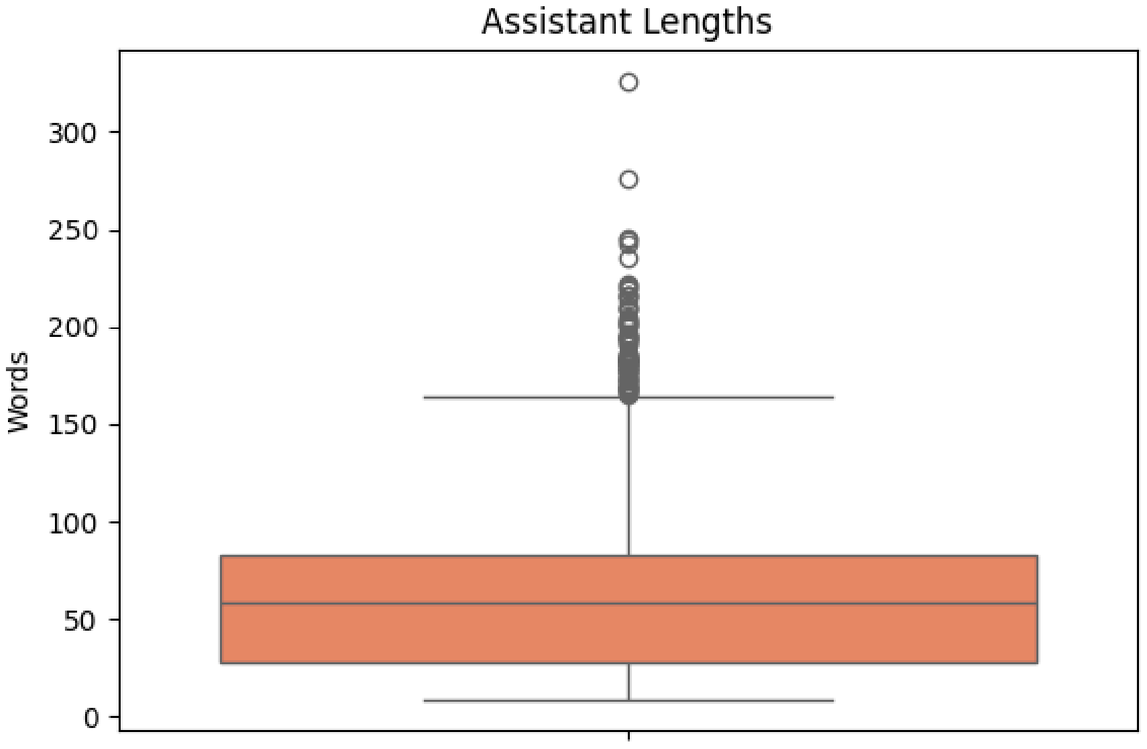
Execution VES

Rouge L



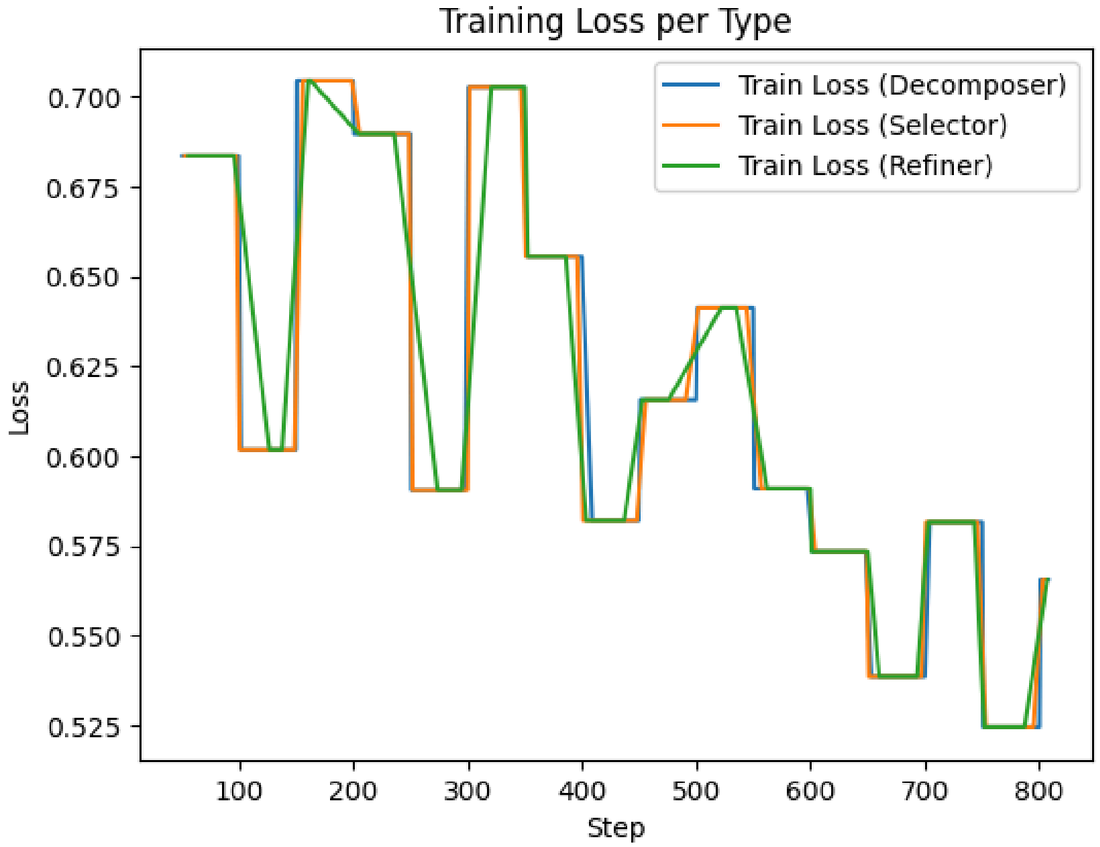
- Decomposer count: 2029
- Selector count: 1009
- Refiner count: 337


User Lengths
Words
1200
1000
800
600
400

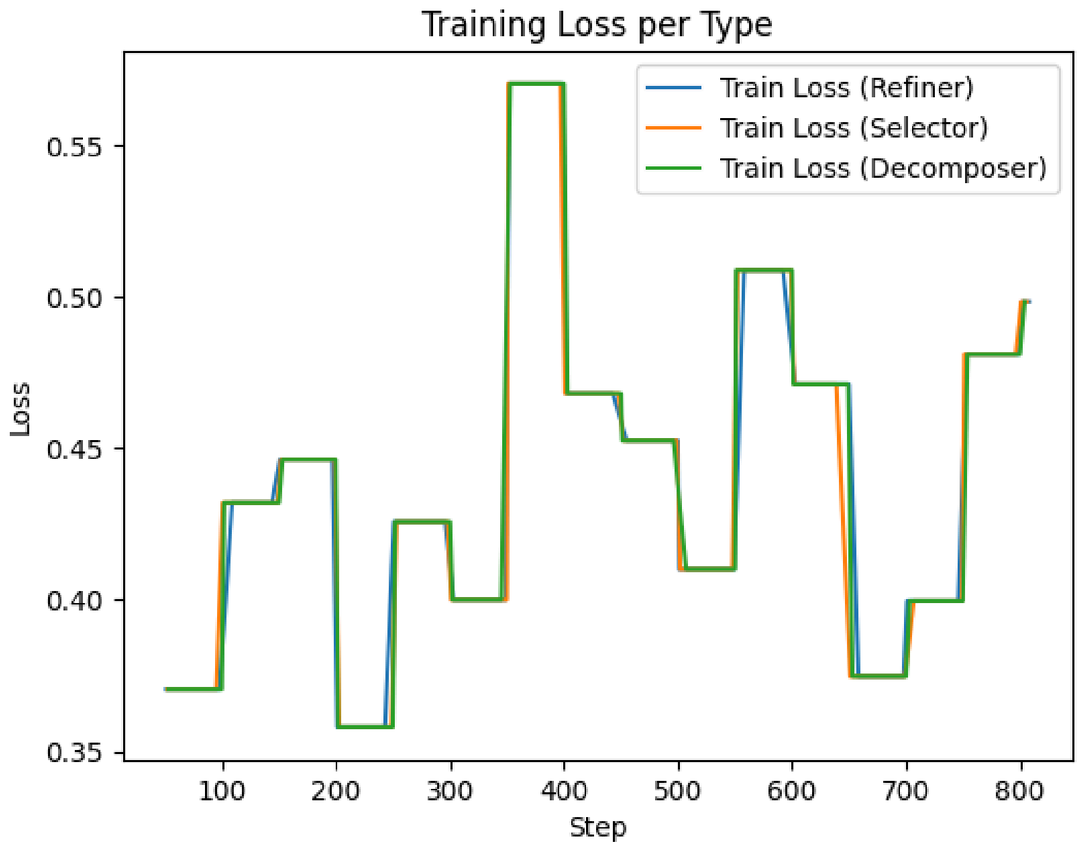

Assistant Lengths
Words
300
250
200
150
100
50
0

Experiment 1: Task balance vs. imbalance

Real distribution

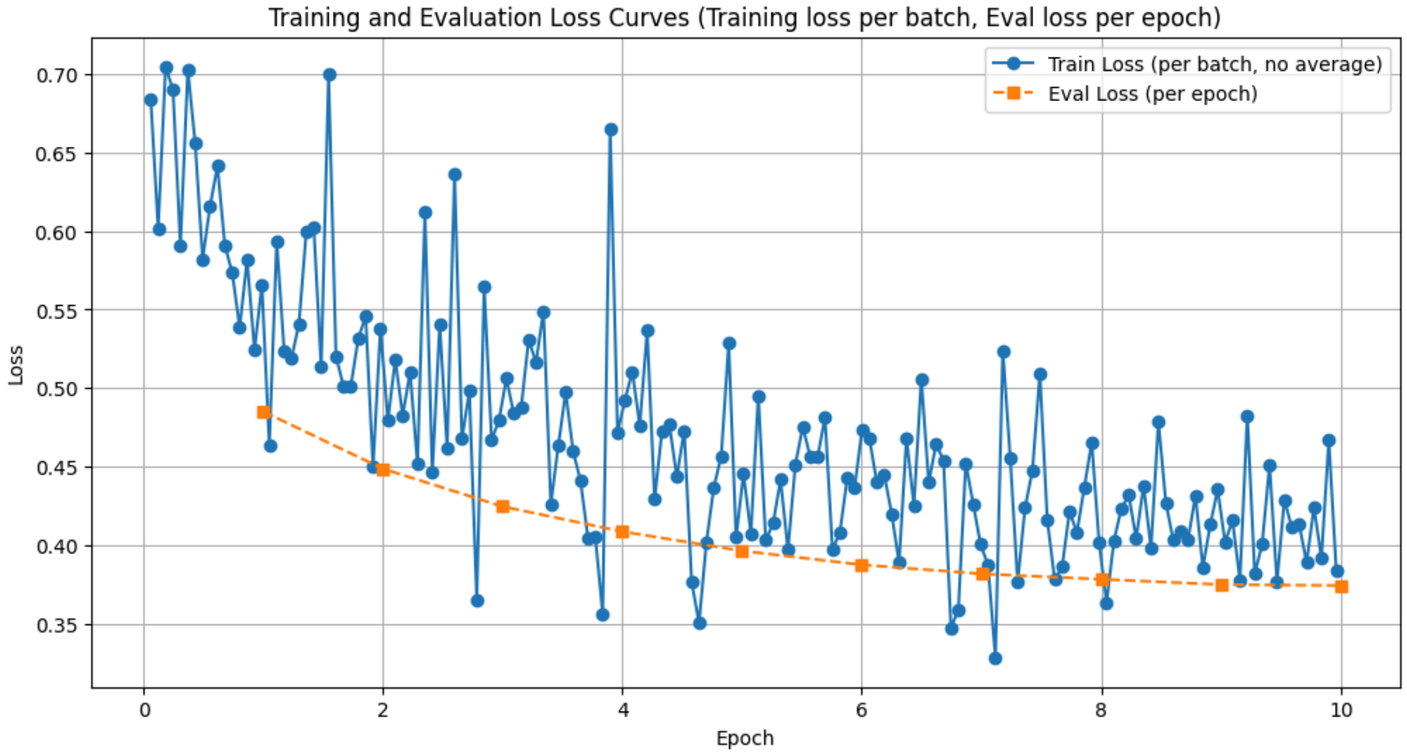


Undersampling



Real distribution

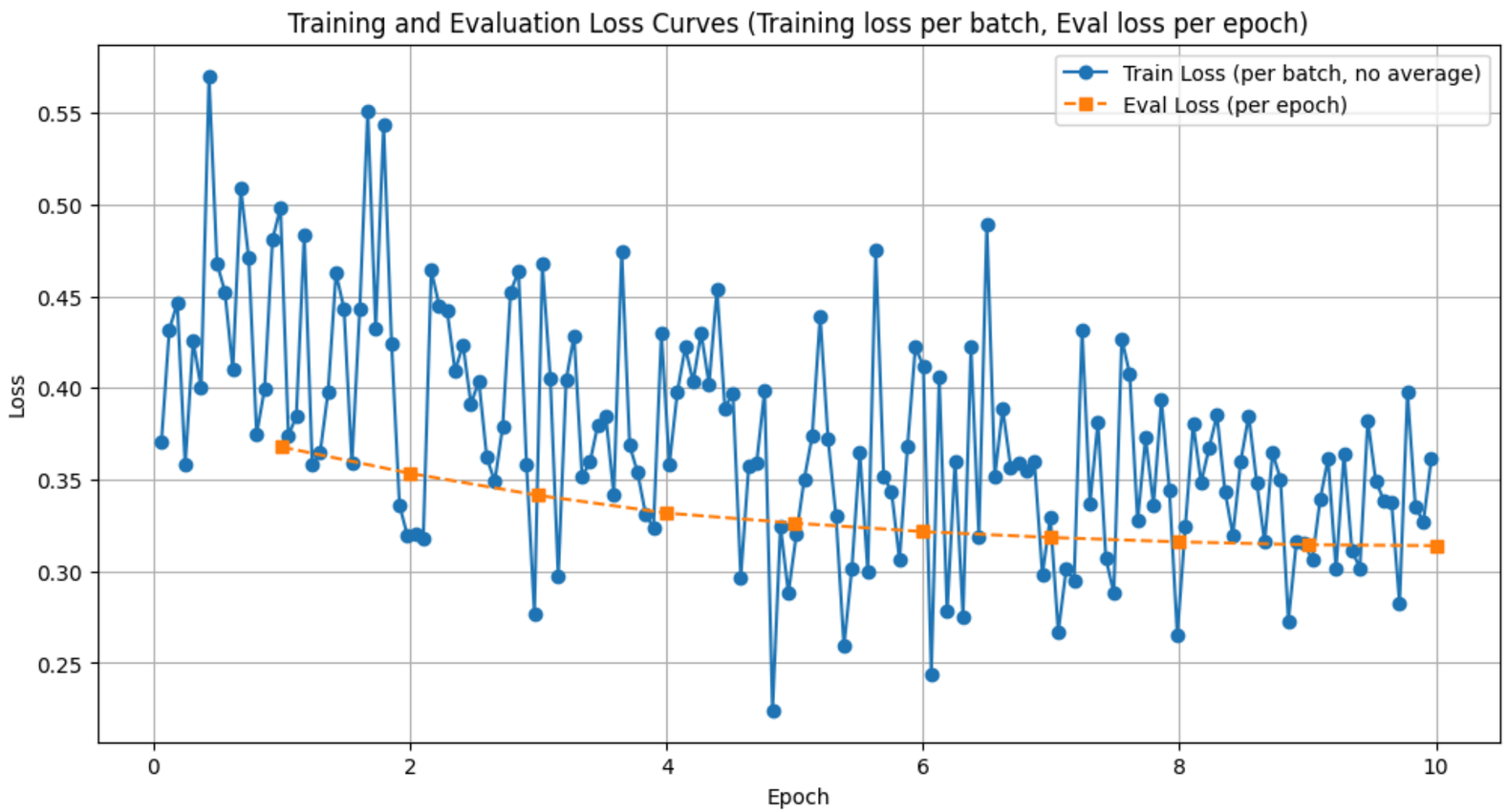
Epoch	Training Loss	Validation Loss
1	0.565600	0.485127
2	0.538100	0.448692
3	0.479700	0.424645
4	0.471500	0.408683
5	0.405000	0.396389
6	0.436800	0.387550
7	0.400500	0.381792
8	0.401500	0.378341
9	0.435400	0.374937
10	0.384100	0.374343



- Rouge_L: 0.354
- EM: 0

Real distribution

Epoch	Training Loss	Validation Loss
1	0.498400	0.367861
2	0.319400	0.353492
3	0.277000	0.341445
4	0.429600	0.331809
5	0.288100	0.326303
6	0.422600	0.321671
7	0.329600	0.318466
8	0.265200	0.316073
9	0.315600	0.314496
10	0.361300	0.313996



- Rouge_L: 0.19
- EM: 0

Experiment 1: Task balance vs. imbalance

Experiment 2: Batch size per device

Experiment 3: Learning rate

1

Complementary

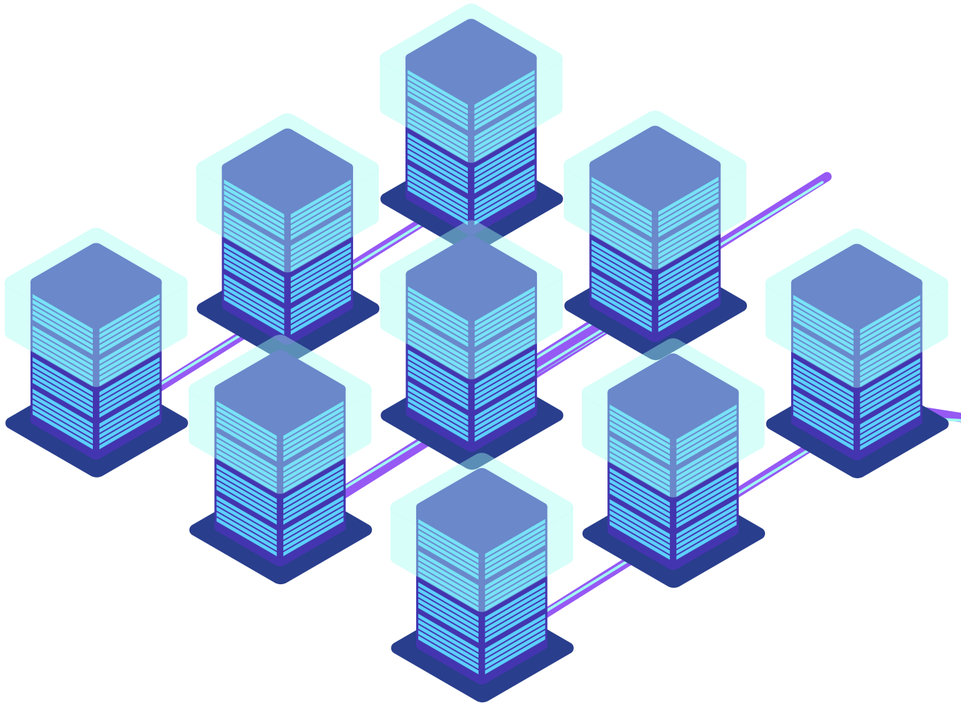
2

Early Stopping
Liger Kernel

3

Same aim

Batch sizes



4

Best pair: 0.227, 0.27
Rouge: 0.36

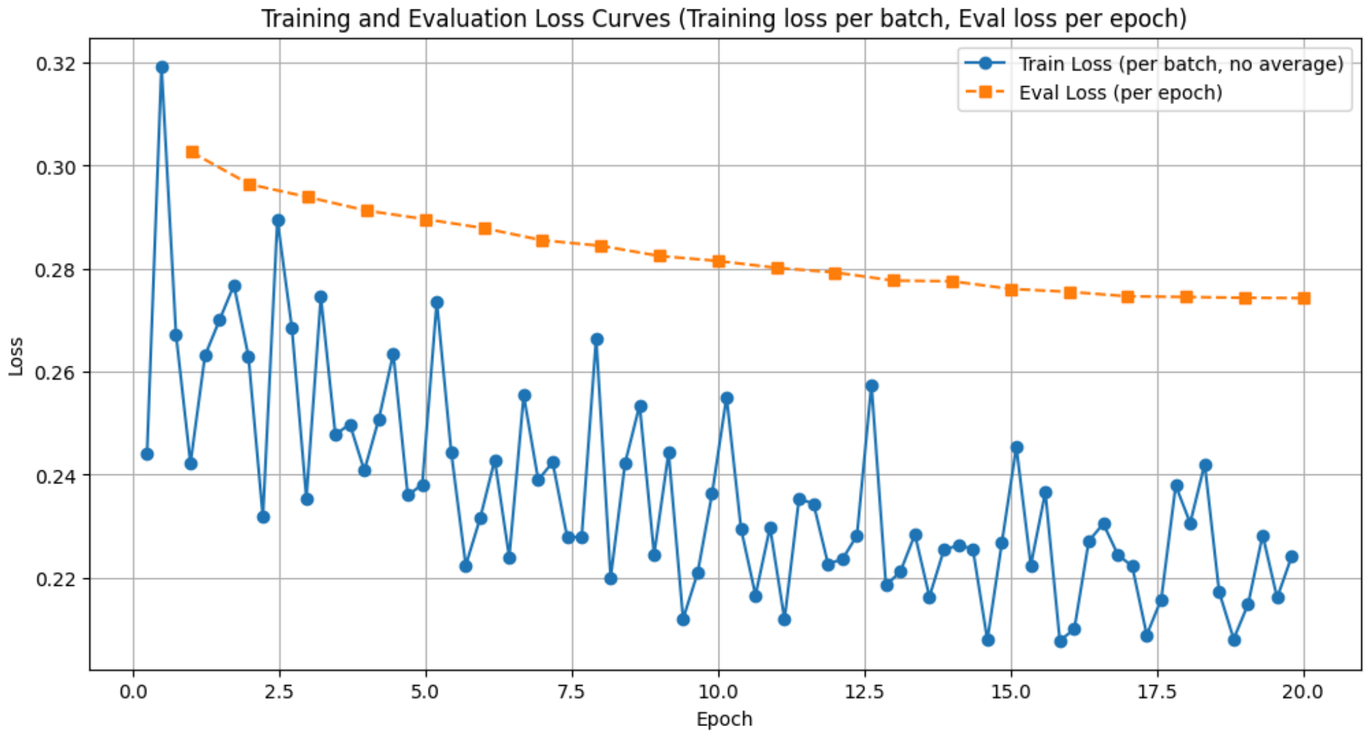
8

Best pair: 0.47, 0.45
Rouge: 0.33

LR: [3e-5, 5e-5, 1e-4]

BS: [4, 8]

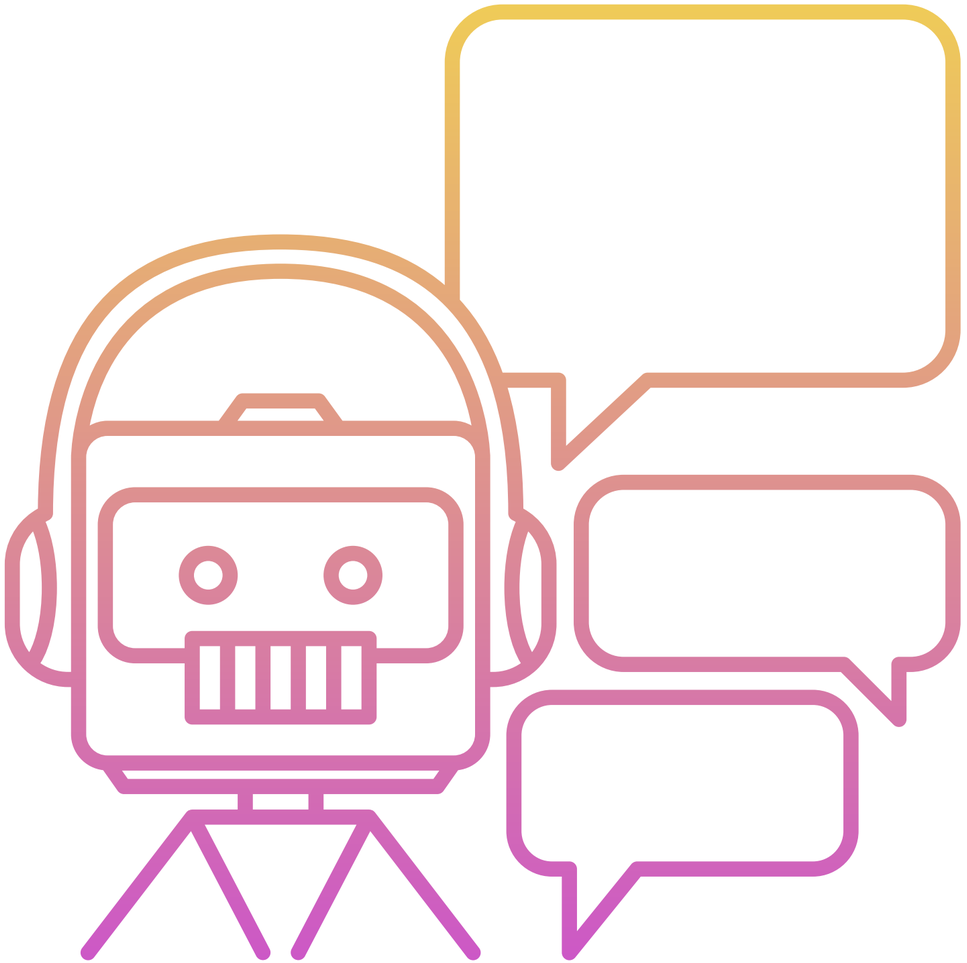
Best pair: 0.2, 0.27
EX: 0.05
Rouge: 0.4



Experiment 4: Generation

Task Type	Correct	Incorrect	Accuracy	Notes
Decomposer	**0**	**2**	**0%**	**Limit tokens, hallucination**
Refiner	**0**	**2**	**0%**	**Misinterpreted schemas**
Selector	**0**	**2**	**0%**	**malformed structure**

Conclusion



It was possible to finetune a very ligthweight model with so few data?


Q
A

[1] Shi, L., Tang, Z., Zhang, N., Zhang, X., & Yang, Z. (2024). A Survey on Employing Large Language Models for Text-to-SQL Tasks. http://arxiv.org/abs/2407.15186
[2] Wiesinger, J., Marlow, P., Vuskovic, V., Huang, E., Xue, E., Sercinoglu, O., Riedel, S., Baveja, S., Gulli, A., Nawalgaria, A., Mollison, G., & Haymaker, J. (2024) Google Whitepaper on AI Agents : Google : Free Download, Borrow, and Streaming : Internet Archive. (n.d.). Retrieved May 4, 2025, from https://archive.org/details/google-ai-agents-whitepaper
[3] Wang, B., Ren, C., Yang, J., Liang, X., Bai, J., Chai, L., Yan, Z., Zhang, Q.-W., Yin, D., Sun, X., & Li, Z. (2023). MAC-SQL: A Multi-Agent Collaborative Framework for Text-to-SQL. http://arxiv.org/abs/2312.11242
[4] Goodfellow, I., Bengio, Y., & Courville, A. (n.d.). *Deep Learning*.
[5] Tunstall, L., von Werra, L., & Wolf, T. (n.d.). *Natural Language Processing with Transformers Building Language Applications with Hugging Face*.