Methodologies for the Interpretation, Evaluation, and Enhancement of Large Language Models

Andris Hohbergs;
Advisor: Dr.sc.comp. Agris Šostaks

Scope of my work and relevance

- ➔ My goal is to identify how **practically useful** LLMs can be for real world use cases. We can see the usefulness in code generation, but other domains still seem to lag behind.

- ➔ The topic is probably too broad, but I have touched upon all of these subjects;

Work completed I: 01.06.25 - 31.08.25

Project at LU “Mākslīgā intelekta metožu piemērotības analīze Eiropas Savienības fondu projektu jomā (VPP-CFLA)”

- ➔ Publication: “Dataset and Baseline Implementation for Retrieval-Augmented Procurement Validation”; Accepted at 14th Computing Conference 2026, London
- ➔ “Retrieval-Augmented Generation for Procurement Validation”; Accepted at ICAART 2026 : 18th International Conference on Agents and Artificial Intelligence; Marbella, Spain

First author publication based on my master’s thesis

- ➔ “Crosscoders for small LLM checkpoint diffing: Qwen and Llama model case study”; Accepted at 14th Computing Conference 2026, London, 9-10 July.

Work completed II: 02.10.25 - 31.08.26 (expected)

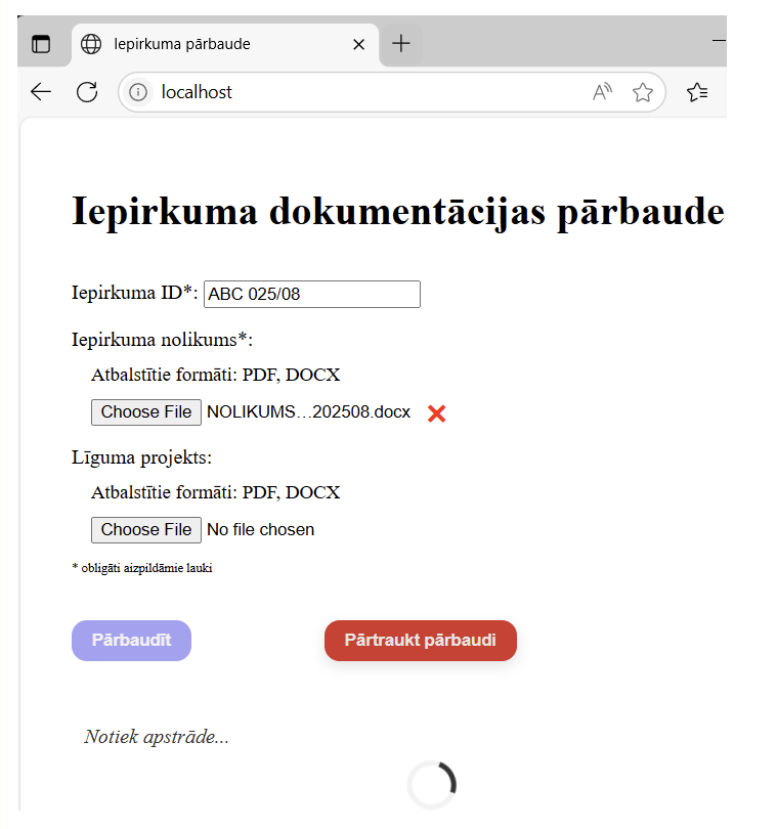
Project at LUMII for LETA: “Lielo valodu modeļos balstīta virtuālā asistenta izstrāde atbilžu izgūšanai no liela apjoma teksta datu masīviem”

- ➔ Publication: “Towards Reliable Evaluation of Legal QA Systems using LLM-as-a-Judge”; Submitted to Baltic DB&IS 2026: 17th International Baltic Conference on Digital Business and Intelligent Systems
- ➔ Planned:
  - ◆ Applying for new projects
  - ◆ Other publications after project

CFLA project I: The problem

CFLA (Central financing & contracting agency) wants to investigate if its possible to evaluate procurement proposals more quickly using RAG.

- ➔ They gave us real procurement evaluation criteria, proposals and their evaluations by experts (30 document examples, each has ~150 questions/answers based on documentation)
- ➔ They expected a system prototype that can use user provided context, relevant laws to fill their evaluation questionnaire



CFLA project II: Data

- ➔ We manually annotated and mapped all the questionnaire answers;
- ➔ Re-wrote some questions for clarification and split up questions that had dependencies;
- ➔ The data was easy to work with because there were only 3 outputs: “yes”, “no”, “n/a”. We also gave the LLM the option to say: “there is not enough context”.

```yaml
- nr: 9
  question0: "Is the procurement contract concluded for health,
  ↪  social and cultural services covered by CPV codes ...?"
  questions:
  - nr: 9.1
    question: "The contracting authority has reserved the
    ↪  opportunity to participate in the procurement in
    ↪  accordance with the procedure of Chapter 10 of the PIL
    ↪  only to those candidates or tenderers who have been
    ↪  granted the status of a social enterprise and who provide
    ↪  services in the aforementioned areas, ..."
    check: "other"
    PIL:
     - chapter: 10
     - chapter: 16
       pt:
        - 2
  - nr: 9.2
    question: "The contracting authority has referred to Chapter
    ↪  16 of the PIL in the notice on social and other special
    ↪  services"
    PIL:
     - chapter: 16
       pt:
        - 3
  - nr: 9.3
    question: "Does the term of the procurement contract not
    ↪  exceed three years?"
```

CFLA project III: Workflow

Later we also performed fine-tuning on the bge-m3 embedding model by using 7083 Latvian procurement projects generating questions about certain segments and then using this as training data (215,971 segments and 430,137 generated questions).

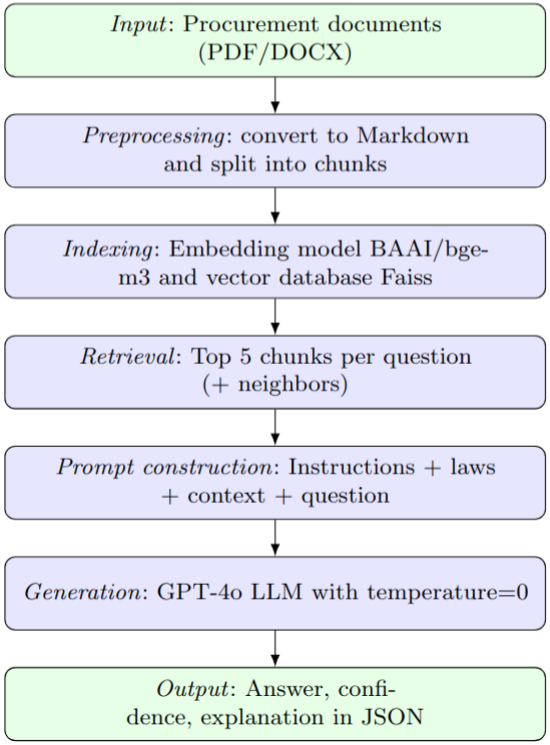


Fig. 4. Workflow of the baseline procurement validation system.

CFLA project IV: Results

Table 4: Accuracies for different embedding models (1 – paraphrase-multilingual-mpnet-base-v2, 2 – bge-m3, 3 – fine-tuned bge-m3).

Model	Overall acc.	Yes/no acc.
1	55.45	74.51
2	66.81	76.91
3	**70.73**	**78.46**

CFLA project V: Conclusions

- ➔ LLMs that were SOTA at the time (GPT-4o) performed quite well on this task when provided the context;
- ➔ We did not do more rigorous analysis to find false positives - i.e., correct answer but incorrect reasoning.
- ➔ Narrowly fine-tuning an embedding model with such a naive method does improve performance;

Crosscoders I: The problem

- ➔ LLMs that are fine-tuned (for example, instruction following models) unlock useful abilities;
  - And sometimes not so useful - sycophancy, misalignment;
- ➔ We don’t know what changes;
- ➔ Studying the differences between models should be easier than studying the whole.

Crosscoders for Small LLM Checkpoint Diffing: Qwen and Llama Model Case Study

Andris Hohbergs[1,2], Uldis Lavrinovičs[3], and Agris Šostaks[1]

[1] Institute of Mathematics and Computer Science, University of Latvia
[2] Faculty of Science and Technology, University of Latvia
[3] LETA, Latvian Information Agency
andris.hohbergs@lumii.lv, uldis.lavrinovics@leta.lv, agris.sostaks@lumii.lv

Abstract. Understanding how large language models (LLMs) internalize and modify their representations during fine-tuning remains a fundamental challenge in interpretability research. Without insight into these changes, models risk exhibiting unpredictable and undesirable behaviors, such as sycophancy or misalignment, which limit their reliability and trustworthiness. This work explores the use of a sparse autoencoder technique - L1 crosscoders - to extract and compare interpretable latent features across base and instruction-tuned LLM variants. We also examine a prior work variant of crosscoders - BatchTopK, which appear to better capture truly model-specific features. Our experiments span several open LLMs (Qwen2.5-0.5B, Qwen2.5-1.5B, and Llama3.2-1B). The results show that model-specific, semantically meaningful features can be extracted even from relatively small models, although L1 crosscoders tend to produce dense, less interpretable latents. If widely adopted, this technique could serve as a standard tool for auditing and aligning LLMs across training checkpoints, fine-tuning objectives, and safety-critical systems.

Keywords: Large language models, Crosscoders, Mechanistic interpretability, Sparse autoencoders, Latent feature extraction, Model comparison

1 Introduction

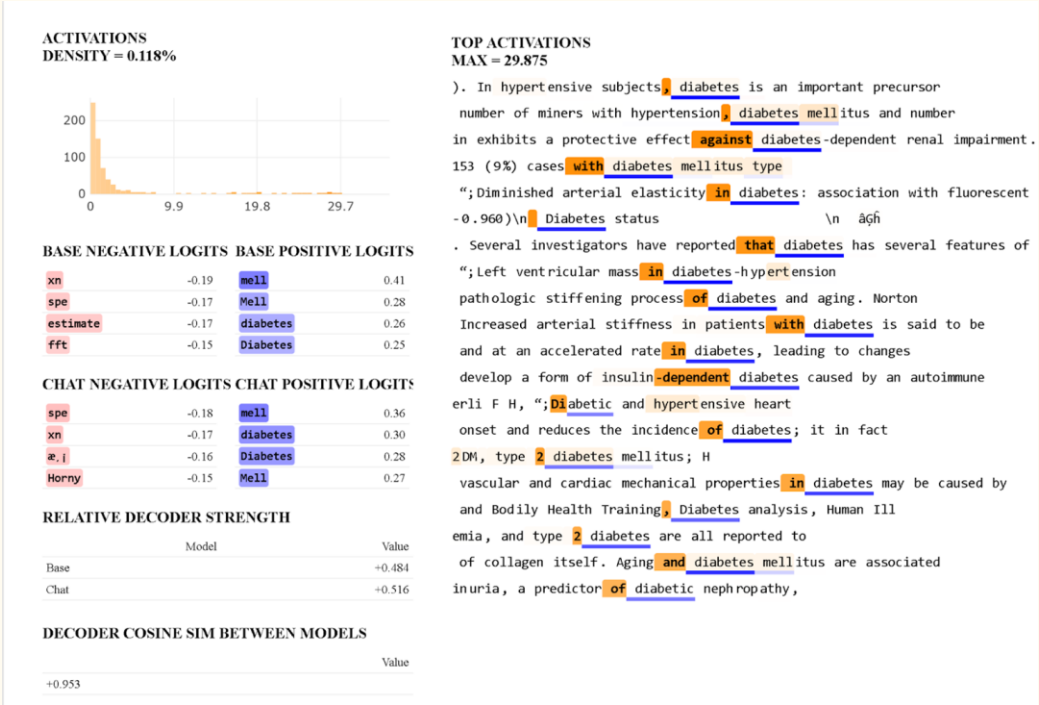
Large language models (LLMs) are increasingly deployed in practical systems, yet their behavior after fine-tuning remains difficult to audit. Standard benchmark-based evaluation is convenient and comparable across models, but it often fails to reveal *how* a model arrives at an answer, is vulnerable to data contamination, and does not reliably predict generalization in real settings. As models are adapted across checkpoints, subtle representational shifts can introduce undesired behaviors (OpenAI, [1], Betley et al., [2]). At the same time, mechanistic interpretability methods that could expose such changes are costly: generating and storing activations at scale requires substantial computational and disk capacity, which limits systematic studies across families and layers.

Crosscoders II: Objective

Based on prior work [Minder, Dumas et.al, and Anthropic research] investigate the crosscoder feature extraction method for small and untested models (Qwen-2.5 0.5, 1.5B, Llama-3.2-1B).

1. See if the method is extendable;

2. Document interpretable features and experiments.

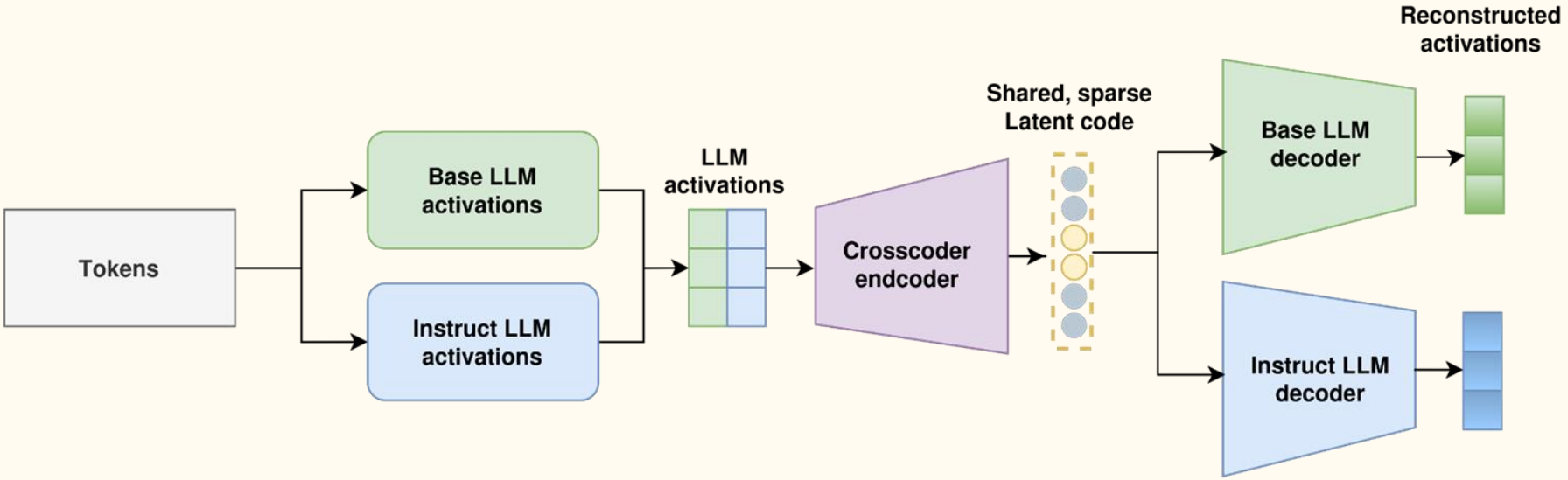


Llama3.2-1B (base) and Llama3.2-1B-instruct crosscoder shared interpretable features example. Feature that activates before outputing “diabetes” token.

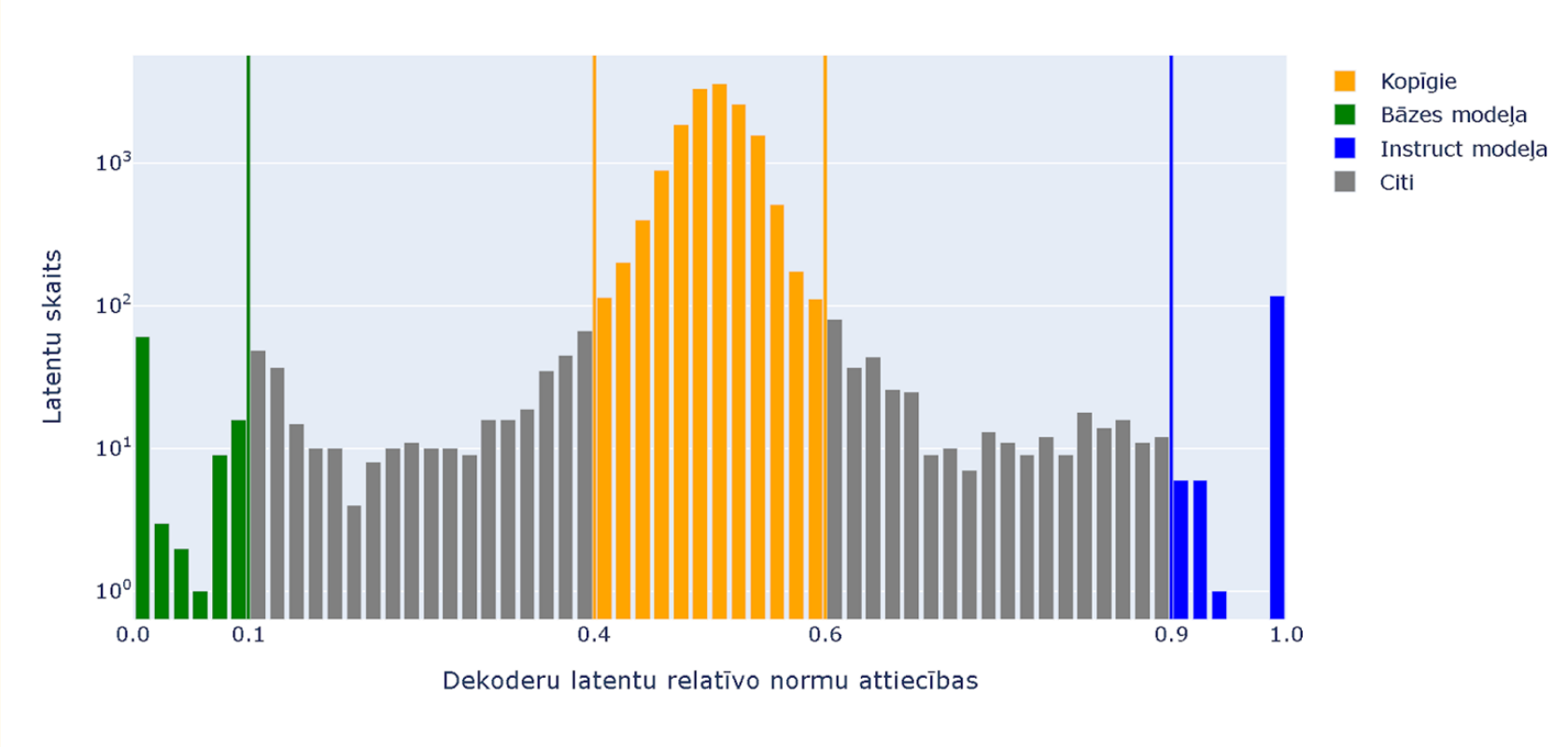
Latenta ID	Latenta tips	Darba autora interpretētā nozīme latentam	Latenta aktivāciju piemēri (jo košāka oranžā krāsa, jo stiprāka aktivācija; treknrakstā stiprākā aktivācija)
12901	*Instruct*	Latents, kas aktivizējas, kad ir runa par dabīgo valodas apstrādi, mākslīgo intelektu vai kad modelis atbild, ka tas ir mākslīgais intelekts.	GPT is a deep **learning** model for natural language processing [...]; Assistant: As **an** AI language model, I [...]
8777	*Instruct*	Ēdiena gatavošanas un jāšanas nodarbības konteksta noteikšanas latents.	Cooking and **experimenting** with new recipes; **doing** riding lessons to play [...]
1988	Bāzes	Mainīgā vērtības nodefinēšanas latents.	5. Let **s** be f(-5).; 5. Let **w** be n(0) [...]
12771	Kopīgs	Latents, kas aktivizējas tieši pirms "diabetes" tekstvienības, jeb "saki *diabetes*" latents.	In hypertensive subjects**,** diabetes is an important precursor [...]
3003	Kopīgs	Vairāku valodu latents, kas aktivizējas uz vārdu "balanced" jeb līdzsvars.	un equ**ilib**rio justo entre [...]; **a** balanced budget and is the [...]
1060	Kopīgs	Kaisles konteksta noteikšanas latents.	to explore **my** own desires and desires [...]; unleashing **your** deepest, darkest desires
3330	Kopīgs	LaTeX matemātikas formāta konteksta noteikšanas iezīme.	),\,\Psi_{**2**}\left(r\right \mathcal S}}_{M} = {{\
56	Kopīgs	Latents, kas aktivizējas pirms skaitļa, kas ir aptuveni 60, jeb "saki aptuveni 60" latents.	70.7)27 **(**62.8)0 [...]; 8--133.5**)**62.6 (56 [...]; /5\ âGȼ 60 (46--71

Results: Llama3.2-1B Crosscoder interpretable feature / latent examples. In total 16,384 latents were extracted - 131 instruction un 92 base model specific. Feature that contains inappropriate language is marked in red.

Crosscoders III: Architecture

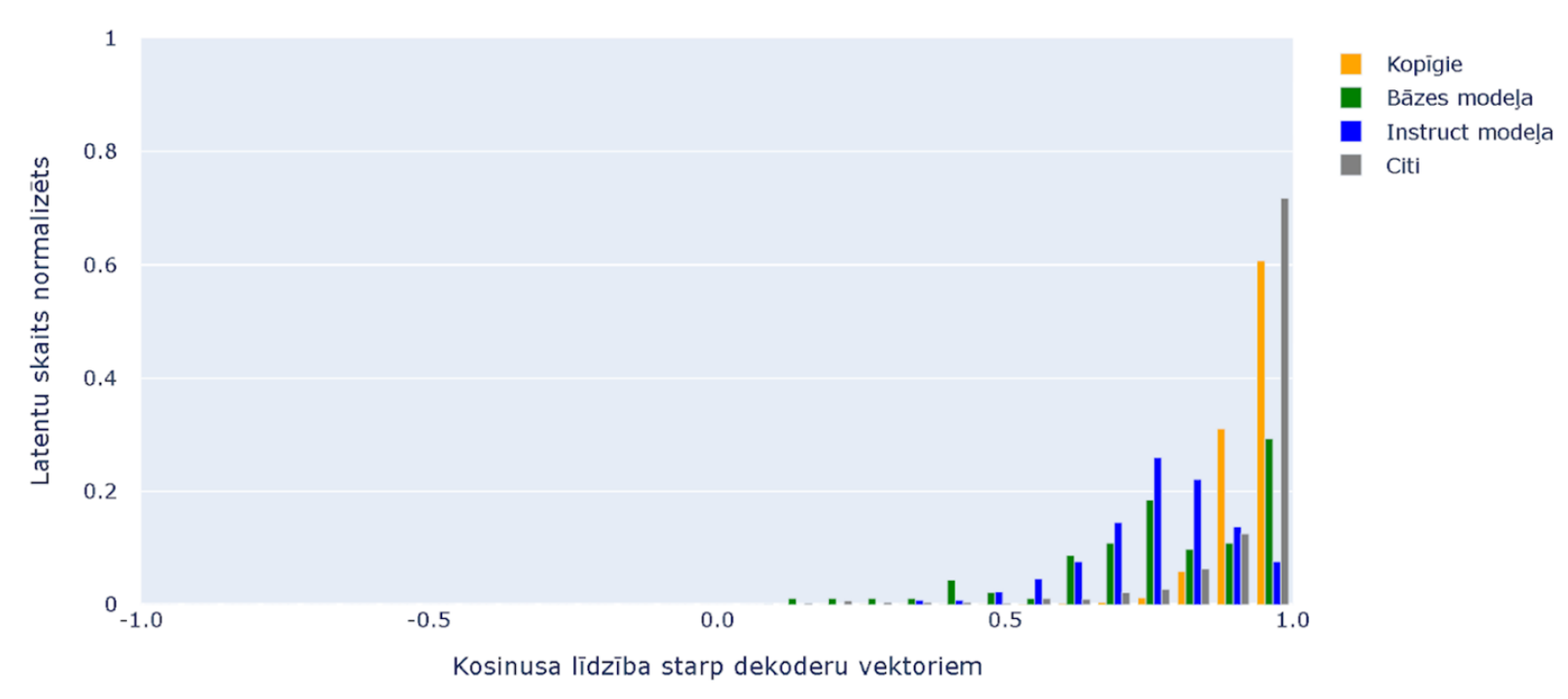


Crosscoders IV: Feature classification



Llama3.2-1B crosscoder decoder latent relative norm (L2) strength.

Crosscoders V: Feature classification



Llama3.2-1B crosscoder decoder latent cosine similarity.

Crosscoders VI: Conclusions

- ➔ LLMs don’t create many new features when fine-tuned, just adjust existing representation;
  - Models can get easily misaligned with fine-tuning.
- ➔ The L1 (initially proposed) Crosscoder does not extract monosemantic model specific features;
- ➔ Features extracted highly depend on the dataset used (no completeness).

Crosscoders VII: Conclusions

Interpretability is costly!

$$2 \times (2\,\text{LLM activations}) \times 896 \times 4\ \ (\text{bytes in FP32}) \times 400{,}000{,}000 \approx 2.61\,\text{TB}$$

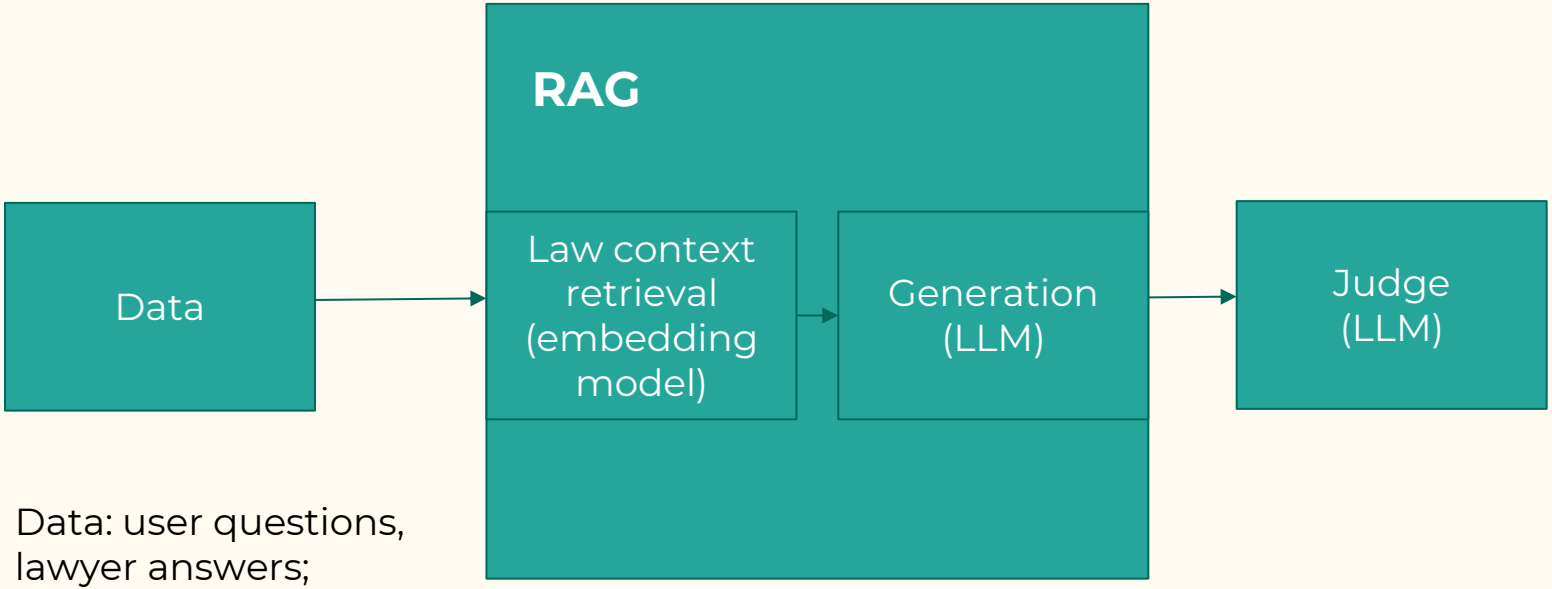
896 = Residual stream for a small Qwen0.5B model typically 2048 to 4096 or even SOT which can be as large as 12k+

LETA project I: The problem

LETA (latvian national information agency) wants to improve it’s services by adding a RAG system that can generate answers to user questions using their legal acts database (legis).

As a proxy for this problem we are using *lvportals* data, which contains real user legal questions, lawyer answers and references to laws.

LETA project II: Pipeline / tuning-dimensions



Data: user questions, lawyer answers; gold_context - references - likumi.lv law points. 31k total samples. 4.5k with only likumi.lv laws.

LETA project III: Q/A example

QUESTION

Situācija ir šāda: slēgtā teritorijā noticis CSN, kura dēļ ir izsaukta Valsts policija. Piemērots administratīvais sods pēc APK panta. Darba devēja pilnvarotā komisija, izvērtējot notikumu, nav ņēmusi vērā visus notikušā faktus (jebkurā notikumā ir iesaistītas divas puses, kurām abām ir ne tikai tiesības, bet arī pienākumi). Vai darba devējs ir tiesīgs piemērot jebkādu papildu sodu, arī disciplinārsodu, naudas sodu u. c., ja iekšējie kārtības noteikumi paredz, ka teritorijas ceļu satiksmes noteikumi nav pretrunā ar Ceļu satiksmes likumu (piemērotie soda punkti ir 4 reizes lielāki nekā soda punkti attiecībā pret CSL paredzētajiem)? Vai šis gadījums ir uzskatāms par dubulto sodīšanu?

RESPONSE

Darba likuma 90. pants nosaka, ka darba devējs par darba kārtības vai darba līguma pārkāpumu var izteikt tikai rakstveida piezīmi vai rājienu [1]. Piezīmi vai rājienu jāizsaka ne vēlāk kā viena mēneša laikā no pārkāpuma atklāšanas, un par katru pārkāpumu var izteikt tikai vienu šādu sankciju [1]. Norma neparedz darba devēja tiesības piemērot papildu disciplinārus vai naudas sodus, kas pārsniedz šo piezīmes/rājiena ietvaru. Tādējādi, ja darba devējs jau izteicis piezīmi vai rājienu par notikumu, papildu sankcijas (diskriminācijas sods, naudas naudas sods u.c.) būtu neatļautas un var tikt uzskatītas par dubulto sodīšanu. No sniegtā normatīvā materiāla izriet, ka darba devējs var piemērot tikai vienu piezīmi vai rājienu, pārējus sankcijas tiesiski nevar piešķirt.

REFERENCE

Saskaņā ar Darba likuma 90. pantu darba devējs par noteiktās darba kārtības vai darba līguma pārkāpšanu darbiniekam var izteikt rakstveida piezīmi vai rājienu, minot tos apstākļus, kas norāda uz pārkāpuma izdarīšanu. Ne Darba likuma normas, ne darba aizsardzību regulējošie normatīvie akti neparedz darba devēja tiesības piemērot naudas sodu par pārkāptajiem uzņēmuma iekšējās kārtības noteikumiem.

RAG CONTEXT (1 ITEMS)

Darba likums (LawID: 26019) CITED

90.pants. Piezīme un rājiens
(1) Par noteiktās darba kārtības vai darba līguma pārkāpšanu darba devējs darbiniekam var izteikt rakstveida piezīmi vai rājienu, minot tos apstākļus, kas norāda uz pārkāpuma izdarīšanu.
(2) Pirms piezīmes vai rājiena izteikšanas darba devējs rakstveidā iepazīstina darbinieku ar viņa izdarītā pārkāpuma būtību un pēc tam pieprasa no viņa rakstveida paskaidrojumu par izdarīto pārkāpumu.
(3) Piezīmi vai rājienu izsaka ne vēlāk kā viena mēneša laikā no pārkāpuma atklāšanas dienas, neieskaitot darbinieka pārejošas darbnespējas laiku, kā arī laiku, kad darbinieks ir atvaļinājumā vai neveic darbu citu attaisnojošu iemeslu dēļ, bet ne vēlāk kā 12 mēnešu laikā no pārkāpuma izdarīšanas dienas. Par katru pārkāpumu var izteikt tikai vienu piezīmi vai rājienu. Darba devējam ir pienākums izsniegt darbiniekam rakstveida rīkojumu, ar kuru darbiniekam ir izteikta piezīme vai rājiens.
(4) Darbiniekam ir tiesības prasīt piezīmes vai rājiena atcelšanu mēneša laikā no piezīmes vai rājiena izteikšanas dienas šā likuma 94.pantā noteiktajā kārtībā. Ja, izskatot sūdzību par piezīmes vai rājiena atcelšanu, darba devējs pieņem lēmumu piezīmi vai rājienu neatcelt, darbiniekam ir tiesības celt prasību tiesā viena mēneša laikā no dienas, kad saņemts attiecīgais darba devēja lēmums. Ja darba devējs šā likuma 94. panta otrajā daļā noteiktajā laikā nav izskatījis sūdzību un nav sniedzis darbiniekam atbildi par pieņemto lēmumu, uzskatāms, ka darba devējs piezīmi vai rājienu ir atcēlis.
(5) Ja gada laikā no piezīmes vai rājiena izteikšanas dienas darbiniekam nav izteikta jauna piezīme vai rājiens, viņš uzskatāms par disciplināri nesodītu.

LETA project IV: Judge and rubrics

We use boolean type questions; Qualitative score from 1-4.

Coverage of Questions and Relevance

- ➔ **answers_all_questions** (mentions all user questions); **Bold** = hard constraint
- ➔ no_unrelated_content

Internal consistency - no internal hallucinations, illogical internal conclusions

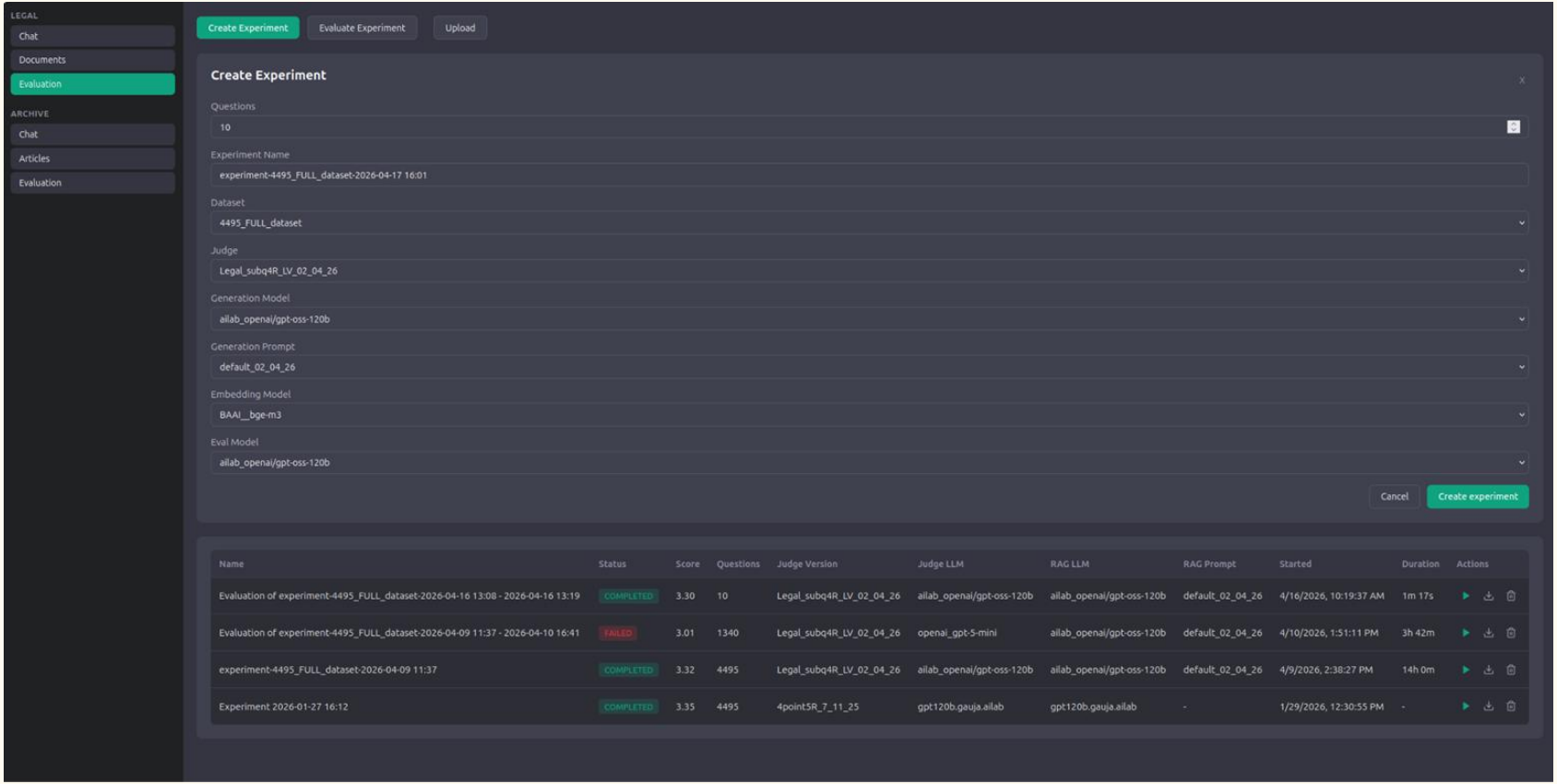
Supported by context

- ➔ **has_citations**
- ➔ **all_relevant_context_used**
- ➔ **correct_utilization_of_context**
- ➔ **no_external_information_used** - no external halucinations
- ➔ correct_formatting_of_references

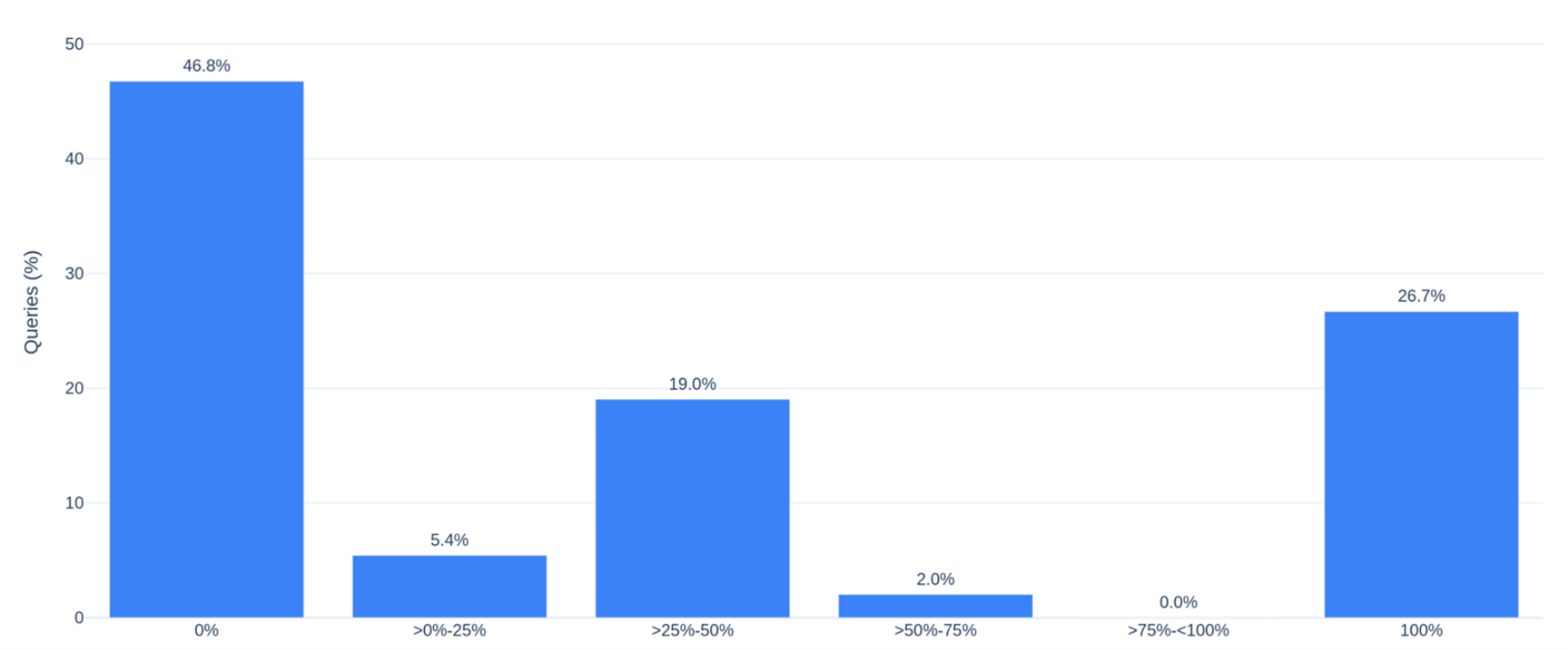
Consistency with reference

- ➔ **no_missing_context_in_gold***
- ➔ **no_conflicting_conclusion**
- ➔ **no_practical_omissions**

LETA project V: Evaluation framework



LETA project VI: Embedding evaluation



Embedding recall histogram BAAI_bge-m3; 588 queries.

LETA project VII: Current problems

- ➔ Data
  - Difficult to work with, contains confusing legal jargon;
  - References are not specific, missing or lawyers make logical leaps based on outside knowledge;
  - Laws change over time.
- ➔ Retrieval
  - Low precision & recall;
  - Relying on LLM for dataset.
- ➔ Generation
  - Not enough computing resources - have to rely on weaker models.
- ➔ Judge
  - Not enough computing resources;
  - We have no legal experts to evaluate our final answers.

LETA project VIII: Conclusions

- ➔ When dealing with a weaker model and larger contexts models seem to struggle less if you use control tokens: <law_context><law_context/>
- ➔ Best embedding model for our use case: BAAI_bge-m3*
- ➔ Best price* to performance LLM (at the moment): GPT-5-mini

- ➔ You should be as specific with your prompts as posible (other error example)
- ➔ Gpt-oss-120B is ok, but it struggles with Latvian language, sometimes still makes logic errors. ~70% accuracy with gold context***; It does not judge itself well
- ➔ Using an English prompt doesn't seem to meaningfully increase performance

Summary

- ➔ 1 Leading author publication that I will be presenting in a conference in London (Interpretability)
- ➔ 2 other publications that are published internationally (Evaluation and Enhancement)
- ➔ 1 publication currently submitted to Baltic Jurnal (Evaluation and Enhancement)
- ➔ 1 more publication planned (Evaluation and Enhancement, Interpretablity?)

Thank you for your attention!

Discussion questions:

- ➔ CFLA project
- ➔ Mechanistic interpretability
- ➔ LETA project + Demo