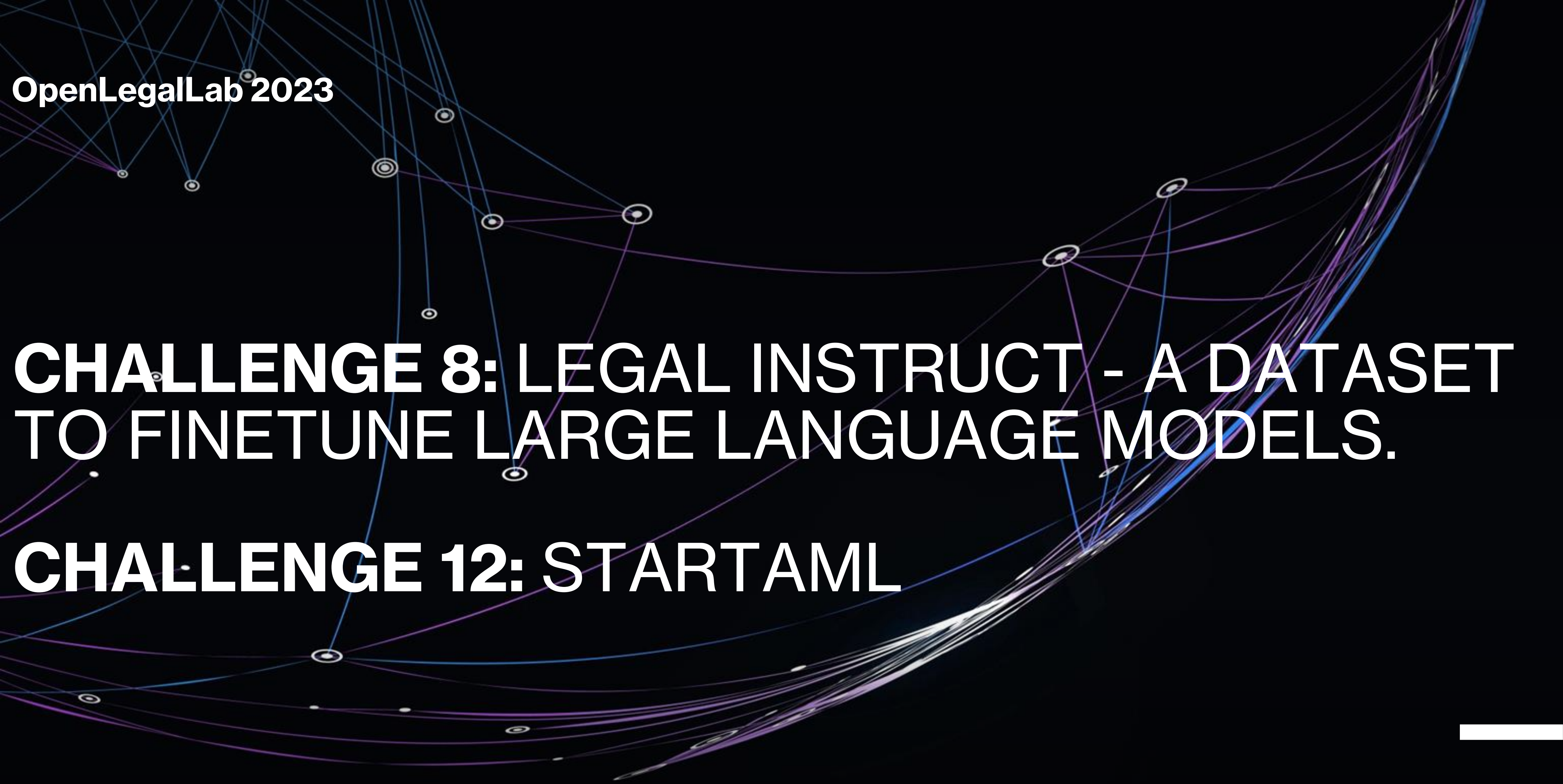


OpenLegalLab 2023

CHALLENGE 8: LEGAL INSTRUCT - A DATASET TO FINETUNE LARGE LANGUAGE MODELS.

CHALLENGE 12: STARTAML

Challenges

- **Challenge 8**: Creation of an openly available Legal-Instruct dataset to finetune large language models for the legal field.
- **Goals:**
  - Create and release a legal-instruct dataset to huggingface with an open license.
  - Finetune a LLM with the created dataset.
- **Challenge 12**: Create a tool that can support counsel in answering legal questions related to the anti-money laundering regulation.

How to create a dataset?

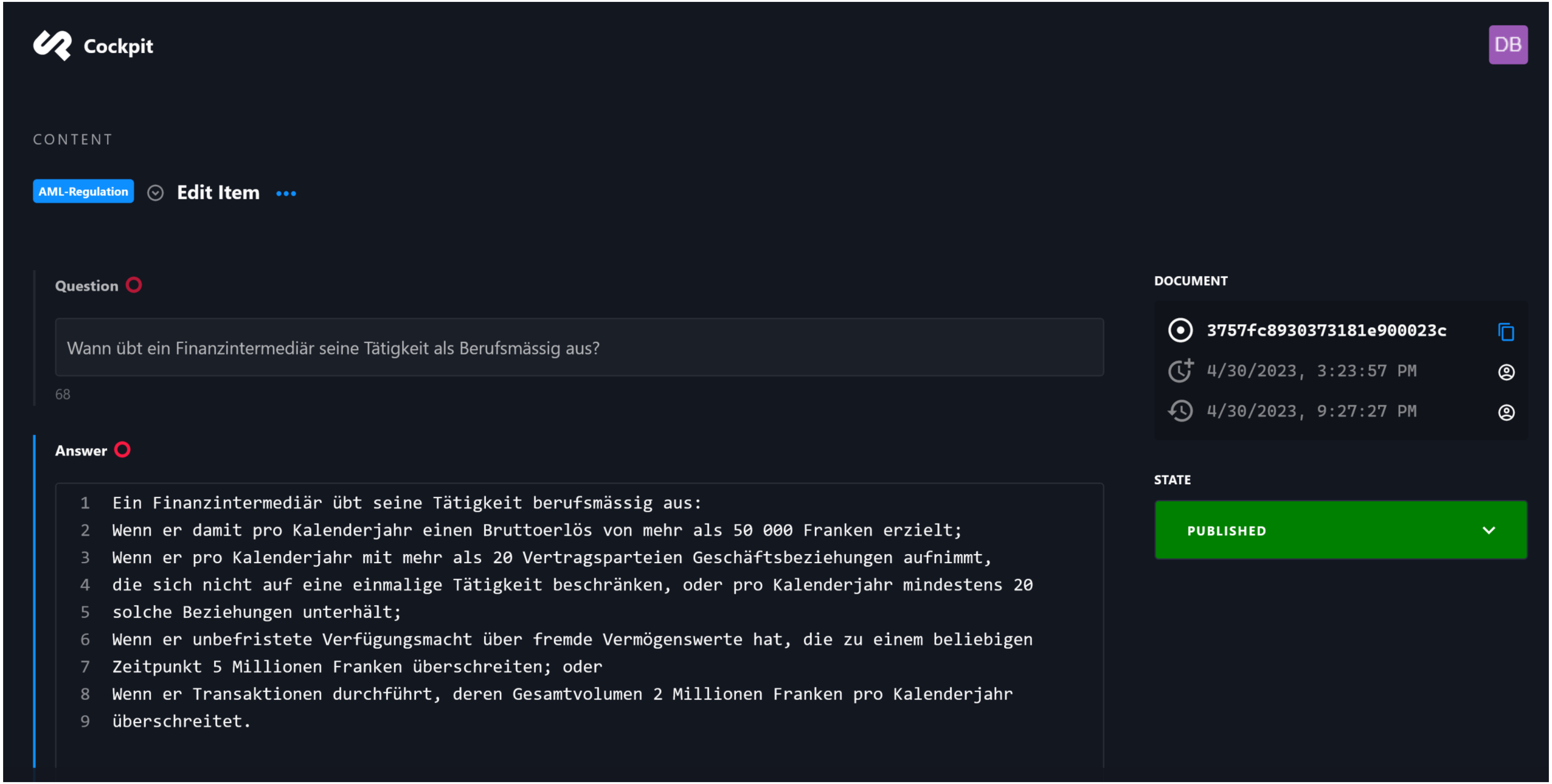
- Manually vs. automated
- Sources: Acts, ordinances, circulars, parliamentary proceedings, private publications such as textbooks, FAQs, etc.
- Are the sources copyrighted? E.g. no university exams, textbooks or AI summaries (e.g. created by openAI).

Manual creation – Anti-money Laundering Regulation

- Manually create questions and answers based on legal texts.
- Sources: AMLA, AMLO, VQF SRO-Regulations, FINMA Circular 2011/1.
- **Result: Created 104 questions and answers with regards to the AMLA regulation.**

Manual creation – FAQs on data and consumer protection

- Source: FAQs on admin.ch:
- Data protection (FR/DE): https://www.bj.admin.ch/bj/de/home/staat/gesetzgebung/datenschutzstaerkung/faq.html
- Consumer protection (FR/DE/EN): https://www.konsum.admin.ch/bfk/de/home/konsumenteninformationen/faq.html
- Manually record questions and answers.
- **Result: Created 60 questions and answers with regards to data and consumer protection.**



Question

Wann übt ein Finanzintermediär seine Tätigkeit als Berufsmässig aus?

68

Answer

Ein Finanzintermediär übt seine Tätigkeit berufsmässig aus:
Wenn er damit pro Kalenderjahr einen Bruttoerlös von mehr als 50 000 Franken erzielt;
Wenn er pro Kalenderjahr mit mehr als 20 Vertragsparteien Geschäftsbeziehungen aufnimmt,
die sich nicht auf eine einmalige Tätigkeit beschränken, oder pro Kalenderjahr mindestens 20
solche Beziehungen unterhält;
Wenn er unbefristete Verfügungsmacht über fremde Vermögenswerte hat, die zu einem beliebigen
Zeitpunkt 5 Millionen Franken überschreiten; oder
Wenn er Transaktionen durchführt, deren Gesamtvolumen 2 Millionen Franken pro Kalenderjahr
überschreitet.

DOCUMENT

3757fc8930373181e900023c
4/30/2023, 3:23:57 PM
4/30/2023, 9:27:27 PM

STATE

PUBLISHED

Automated creation – Federal court decisions (BGE)

Scraped 17’000 federal court decisions at www.entscheidsuche.ch through the elastic API.

Automatic extraction of the following: 1. “Regeste”, 2. Text of decision and 3. Language (DE/FR/IT)

Creation of a dataset as follows:

- Instruction: “Write me a “Regeste”
- Input: Text of the decision
- Output: Regeste

Problems: Most decisions surpass context length (token count) of 2048.

Solution: Take the 25% of the decisions with the lowest token count. Of this set, randomly delete words until no decision surpasses a token count of 1’500 (since court decisions always surpass token estimates).

Result: A dataset of ca. 4’250 leading federal court decisions in DE, FR and IT was created.

Automated creation – Federal court decisions (BGE)

Using federal court decisions from the Research Center for Digital Sustainability dataset (rcds/swiss_law_area_prediction) on Huggingface.co.

Automatic generation keywords through removal of stopwords, lemmatizing the text and randomly removing 20% of the words (DE/FR)

Creation of a dataset as follows:

- Instruction: “Write a full text using the following keywords”
- Input: Keywords
- Output: Initial Text

Problems: Most decisions surpass context length (token count) of 2048.

Solution: Make smaller chunks of a limited word count (approximation to tokens)

Result: A dataset based on ca. 5’000 leading federal court decisions (DE/FR) was created.

Automated creation – Parliamentary proceedings

- Scraped parliamentary proceedings (questions to federal council including answer by federal council) from “Curia Vista” through the API at ws-old.parlament.ch.
- First Problem: API limit with heavy throttling. Thus, only proceedings of last three years were downloaded.
- Proceedings including legal questions were filtered out (legal proceeding if text contained a reference to an article).
- Remaining 1’800 proceedings (600 different proceedings, each in DE/FR/IT).
- Second problem: Most proceedings surpass context length (token count) of 2048.
- Solution: Filtering of decisions to proceedings not surpassing context length of 1’500.
- **Result: A dataset based on 1’388 parliamentary proceedings in DE, FR and IT was created.**

Automated creation – KESB Winterthur

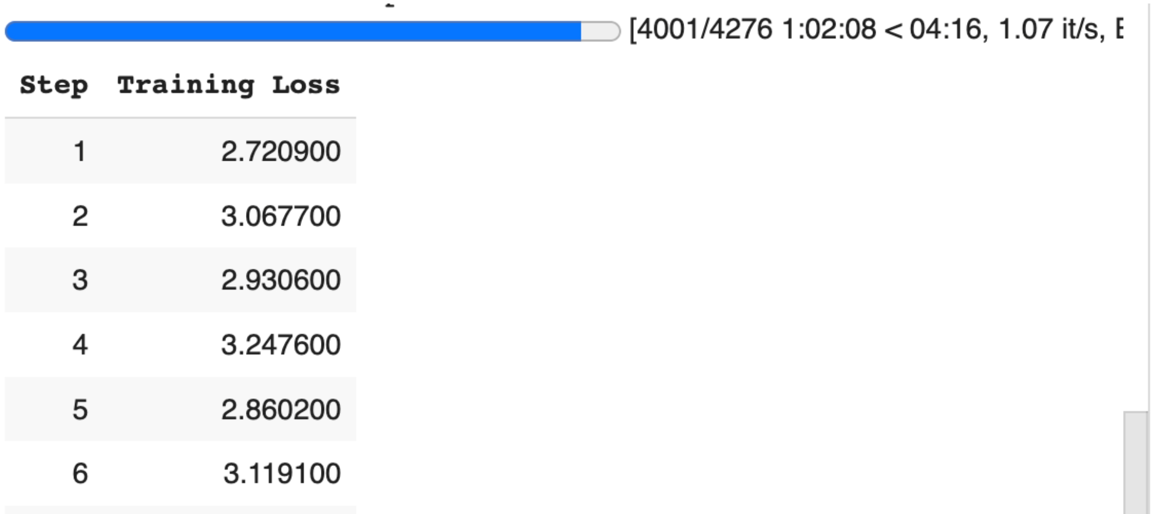
- Scraped FAQs from KESB Winterthur (https://kesb-wa.ch/haeufige-fragen/).
- Automatic creation of questions (input) and answer (output).
- **Result: A dataset based on ca. 100 questions and answers was created.**

First goal: Publication of dataset

- The dataset was released to huggingface with an open license under the following link: https://huggingface.co/oll-2023

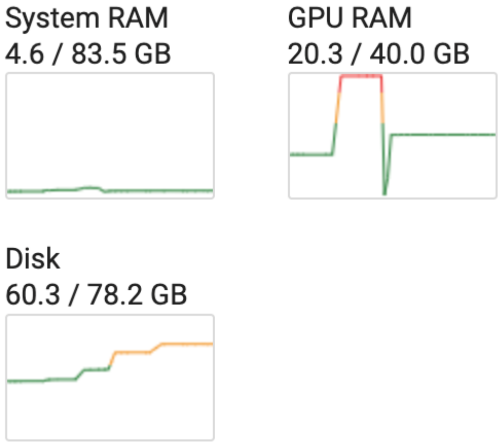
Second goal: Finetuning

- Model: Dolly 2.0 (3B) by databricks
- Finetuning using PEFT (huggingface) and LoRA (Hu/Shen et al.) using our Alpaca-style dataset.



[4001/4276 1:02:08 < 04:16, 1.07 it/s, E

Step	Training Loss
1	2.720900
2	3.067700
3	2.930600
4	3.247600
5	2.860200
6	3.119100

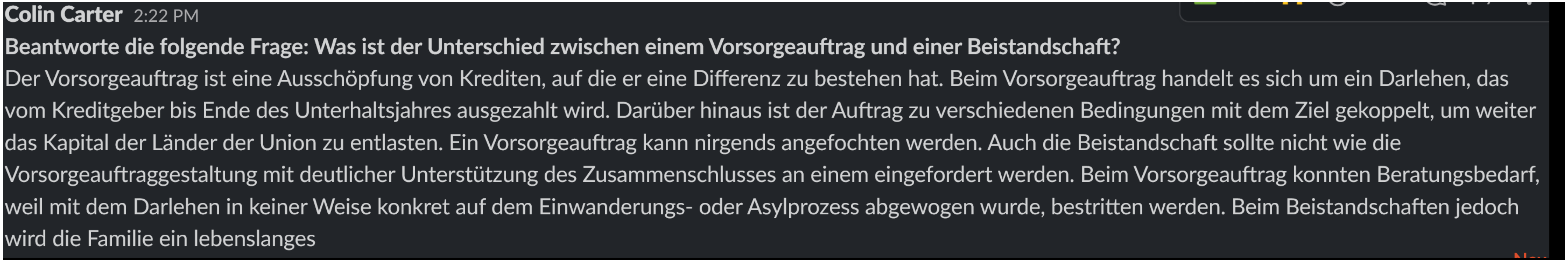


System RAM
4.6 / 83.5 GB

GPU RAM
20.3 / 40.0 GB

Disk
60.3 / 78.2 GB

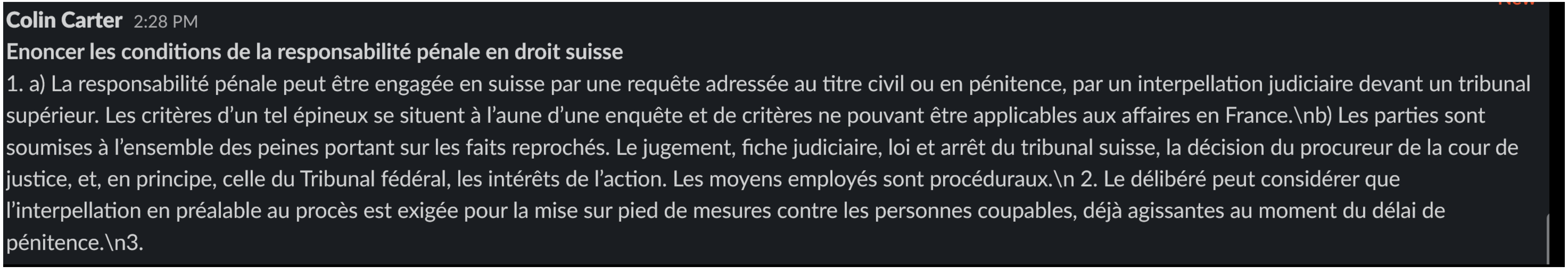
Results



Colin Carter 2:22 PM

Beantworte die folgende Frage: Was ist der Unterschied zwischen einem Vorsorgeauftrag und einer Beistandschaft?

Der Vorsorgeauftrag ist eine Ausschöpfung von Krediten, auf die er eine Differenz zu bestehen hat. Beim Vorsorgeauftrag handelt es sich um ein Darlehen, das vom Kreditgeber bis Ende des Unterhaltsjahres ausgezahlt wird. Darüber hinaus ist der Auftrag zu verschiedenen Bedingungen mit dem Ziel gekoppelt, um weiter das Kapital der Länder der Union zu entlasten. Ein Vorsorgeauftrag kann nirgends angefochten werden. Auch die Beistandschaft sollte nicht wie die Vorsorgeauftraggestaltung mit deutlicher Unterstützung des Zusammenschlusses an einem eingefordert werden. Beim Vorsorgeauftrag konnten Beratungsbedarf, weil mit dem Darlehen in keiner Weise konkret auf dem Einwanderungs- oder Asylprozess abgewogen wurde, bestritten werden. Beim Beistandschaften jedoch wird die Familie ein lebenslanges



Colin Carter 2:28 PM

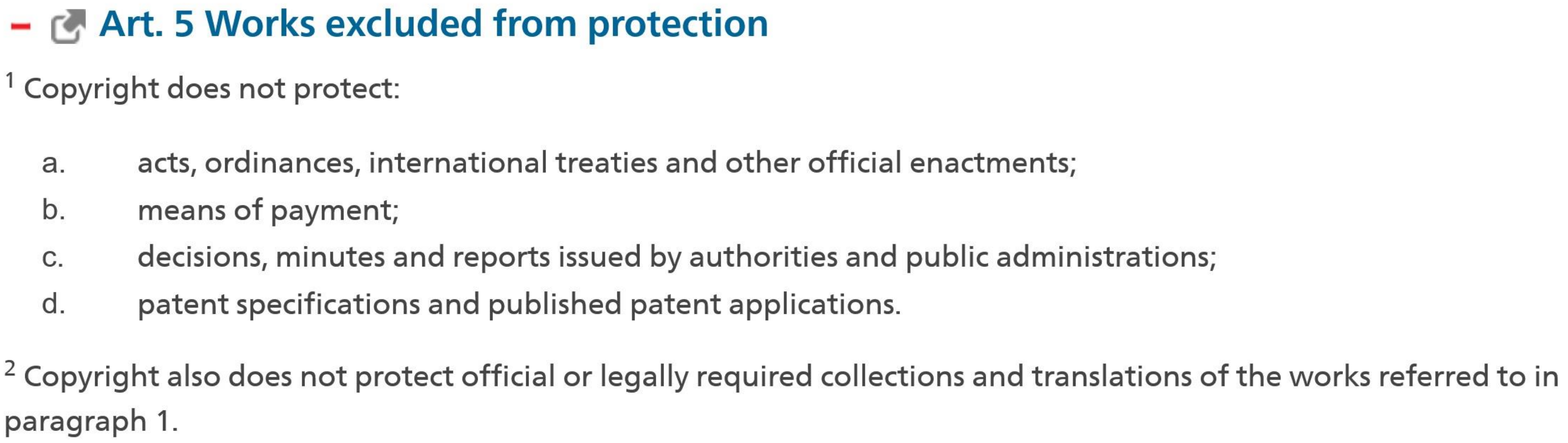
Enoncer les conditions de la responsabilité pénale en droit suisse

1. a) La responsabilité pénale peut être engagée en suisse par une requête adressée au titre civil ou en pénitence, par un interpellation judiciaire devant un tribunal supérieur. Les critères d'un tel épineux se situent à l'aune d'une enquête et de critères ne pouvant être applicables aux affaires en France.\nb) Les parties sont soumises à l'ensemble des peines portant sur les faits reprochés. Le jugement, fiche judiciaire, loi et arrêt du tribunal suisse, la décision du procureur de la cour de justice, et, en principe, celle du Tribunal fédéral, les intérêts de l'action. Les moyens employés sont procéduraux.\n 2. Le délibéré peut considérer que l'interpellation en préalable au procès est exigée pour la mise sur pied de mesures contre les personnes coupables, déjà agissantes au moment du délai de pénitence.\n3.

One take away:
We need common info. assets (1)

- **DILEMMA**
  - **Identification difficult**: Many informational sources are or may be copyrighted – Identification difficult despite art. 5 of the Copyright Act.
  - **Terms & Conditions**: Terms and conditions amongst public institutions (federal and/or cantonal) are not unified, unclear and sometimes do not allow the usage (modification) of the data.
  - **Format**: Information format not unified and not always usable for machine learning (e.g. federal acts are available in XML whereas parliamentary deliberations of the canton of Geneva are only available in PDF).

Federal Act on Copyright and Related Rights

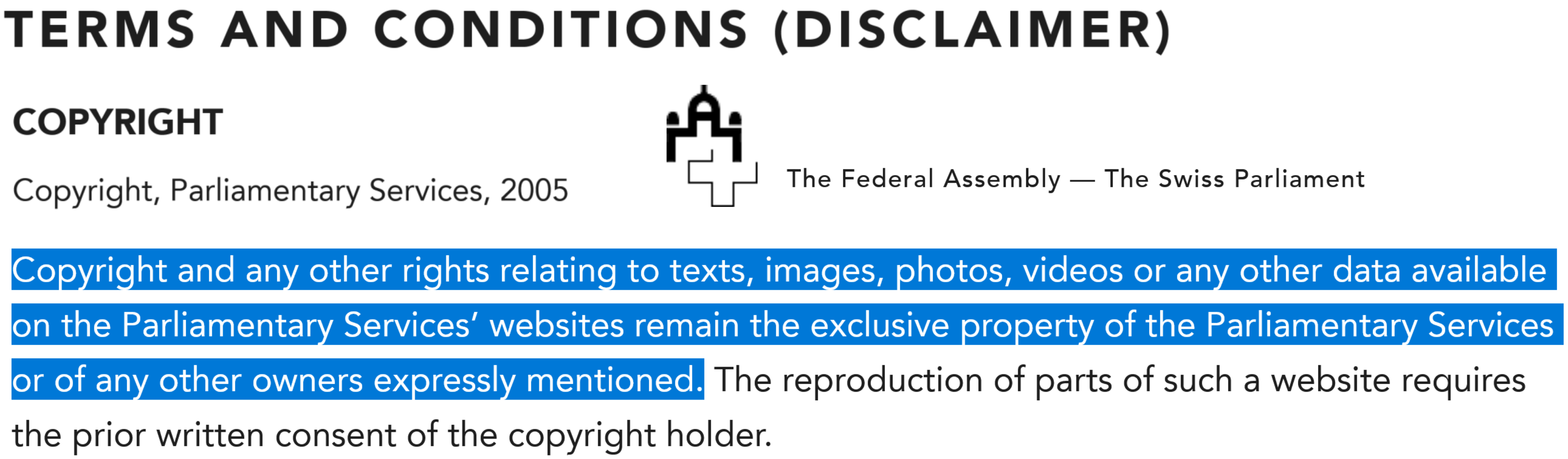


Art. 5 Works excluded from protection

[1] Copyright does not protect:

- a. acts, ordinances, international treaties and other official enactments;
- b. means of payment;
- c. decisions, minutes and reports issued by authorities and public administrations;
- d. patent specifications and published patent applications.

[2] Copyright also does not protect official or legally required collections and translations of the works referred to in paragraph 1.



TERMS AND CONDITIONS (DISCLAIMER)

COPYRIGHT

Copyright, Parliamentary Services, 2005

The Federal Assembly — The Swiss Parliament

Copyright and any other rights relating to texts, images, photos, videos or any other data available on the Parliamentary Services' websites remain the exclusive property of the Parliamentary Services or of any other owners expressly mentioned. The reproduction of parts of such a website requires the prior written consent of the copyright holder.

One take away:
We need common info. assets (2)

- **SUGGESTIONS**
  - *Legally*: Enable open access to all information, i.e. by unifying and amending Terms & Conditions.
  - *Technically*: Use open standard formats across the different websites, such as XML.

Team

- Colin Carter
- Michael Wechner
- Olga Opanasenko
- Evelina Mitrega
- Roger Meier
- Erhard Marro
- Thomas Avanzi
- Quentin Adler
- Naemi Weinmann
- David Billeter

References

- LoRA: Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. ‘LoRA: Low-Rank Adaptation of Large Language Models’. *ArXiv [Cs.CL]*, 2021. arXiv. http://arxiv.org/abs/2106.09685.
- Alpaca: Rohan Taori* and Ishaan Gulrajani* and Tianyi Zhang* and Yann Dubois* and Xuechen Li* and Carlos Guestrin and Percy Liang and Tatsunori B. Hashimot. https://crfm.stanford.edu/2023/03/13/alpaca.html.