Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

RuTaR—A Dataset in Russian for Reasoning about Taxes

Andrey Alibekov
NTR Labs, Tomsk, Russia
Higher IT School, Tomsk State University, Tomsk, Russia
aalibekov@ntr.ai

Aleksandr Migal
NTR Labs, Moscow, Russia
HSE, Saint-Petersburg, Russia
amigal@ntr.ai

Andrey Matenkov
NTR Labs, Moscow, Russia
amatenkov@ntr.ai

Andrey Muryshev
NTR Labs, Moscow, Russia
amuryshev@ntr.ai

Vladislav Bolshakov
NTR Labs, Moscow, Russia
BMSTU, Moscow, Russia
vbolshakov@ntr.ai

Aleksander Kozachenko
NTR Labs, Tomsk, Russia
akozachenko@ntr.ai

Gyuli Mukhtarova
NTR Labs, Moscow, Russia
gmukhtatova@ntr.ai

Nikolay Mikhaylovskiy
NTR Labs, Moscow, Russia
Higher IT School, Tomsk State University, Tomsk, Russia
nickm@ntr.ai

Abstract

In 2024, reasoning have emerged as a new frontier for artificial intelligence and computational linguistics. Reasoning models are typically evaluated either on STEM-related datasets, or on synthetic datasets. This ignores a huge area of human thought—namely, humanitarian. To bridge this gap partially, we present a new open dataset, RuTaR (Russian Tax Reasoning). The dataset consists of modestly modified content of 199 select Ministry of Finances of Russia and Russian Federal Tax Service letters that typically reason to answer some taxpayer question. Despite apparent simplicity of yes/no questions, both off-the-shelf Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) systems struggle to achieve high results on the dataset, with top RAG system studied achieving 77% accuracy.

Keywords: reasoning; dataset; artificial intelligence
DOI: 10.28995/2075-7182-2025-23-XX-XX

RuTaR — набор данных рассуждений о налогах на русском языке

Андрей Алибеков
ООО «НТР Томск», Томск, Россия
Высшая ИТ-Школа Томского Государственного Университета, Томск, Россия
aalibekov@ntr.ai

Александр Мигаль
ООО «НТР», Москва, Россия
НИУ ВШЭ, Санкт-Петербург, Russia
amigal@ntr.ai

Андрей Матенков
ООО «НТР», Москва, Россия
amatenkov@ntr.ai

Андрей Мурышев
ООО «НТР», Москва, Россия
amuryshev@ntr.ai

Владислав Большаков
ООО «НТР», Москва, Россия
МГТУ им. Баумана, Москва, Россия
vbolshakov@ntr.ai

Александр Козаченко
ООО «НТР Томск», Томск, Россия
akozachenko@ntr.ai

Гюли Мухтарова
ООО «НТР», Москва, Россия
gmukhtatova@ntr.ai

Николай Михайловский
ООО «НТР», Москва, Россия
Высшая ИТ-Школа Томского
Государственного Университета,
Томск, Россия
nickm@ntr.ai

Аннотация

В 2024 году рассуждения стали ключевой точкой развития искусственного интеллекта и вычислительной лингвистики. Модели рассуждений обычно оцениваются либо на наборах данных, связанных с точными науками, либо на синтетических наборах данных. Это игнорирует огромную область человеческой мысли — гуманитарную. Чтобы частично восполнить этот пробел, представлен новый открытый набор данных, RuTaR (Russian Tax Reasoning). Набор данных состоит из слегка измененного контента 199 отобранных писем Министерства финансов России и Федеральной налоговой службы, которые с рассуждениями отвечают на вопросы налогоплательщиков. Несмотря на кажущуюся простоту бинарных вопросов, как стандартные большие языковые модели, так и системы генерации на основе поиска испытывают трудности с достижением высоких результатов на наборе данных, при этом лучшая из оцененных система достигает точности лишь в 77 %.

Ключевые слова: рассуждения; набор данных; искусственный интеллект

1 Introduction

In 2024, reasoning have emerged as a new frontier for artificial intelligence and computational linguistics [1-12]. The pivotal productization of o1 reasoning system by Open AI [12] have sparked a wave of research and replication efforts [14-21]. These reasoning models are typically evaluated either on STEM datasets [22-34] (which from our perspective include logics-specific datasets [35-44]), or on synthetic datasets [2, 45-48]. Exceptions are few [49-52]. This ignores the modes of reasoning typical of the humanities, including law and other similar reasoning-intensive fields.

In Russian, there are few original datasets specifically targeted at multi-turn reasoning. Russian SuperGLUE [52] was a fantastic effort for its time, but it is now largely solved [46], including the most relevant to our interests MuSeRC and RuCoS [53]. MultiQ from TAPE [46] is a synthetically constructed 2-hop reasoning benchmark. Therefore, we would like to have a dataset that will be challenging in 2025 and thus require multi-turn reasoning over a large natural corpus.

To address this twofold deficiency, we present a new open dataset, RuTaR (Russian Tax Reasoning). The dataset consists of modestly modified content of select Ministry of Finances of Russia and Federal Tax Service letters that typically answer some taxpayer question. The dataset is made available at https://github.com/rutar-anonymous/RuTaR under the terms of the Creative Commons Attribution License (CC-BY).

Figure 1 provides an example of an original letter of Russian Federal Tax Service. Figure 2 provides examples of data inferred from the same letter: inferred question that the letter answers, inferred short-form answer that can be used to automatically evaluate the reasoning systems and the legal references that can be used to evaluate retrieval systems.

The dataset comprises two distinct types of official letters, 131 of which were issued by the Ministry of Finance and 78 by the Federal Tax Service. Each letter is categorized into one of two possible response classes, with 120 records labeled as "0" and 89 as "1". Following a common binary classification logic, responses labeled as "0" correspond to a summarized answer of "No", while those labeled as "1" correspond to "Yes". On average, each letter contains 4.71 references to legal documents. It's also worth noting that two of the letters do not include any references at all.

Федеральная налоговая служба, рассмотрев обращение <...>, сообщает, что в соответствии с пунктом 3 статьи 29 Налогового кодекса Российской Федерации (далее – Кодекс) уполномоченный представитель налогоплательщика-организации осуществляет свои полномочия на основании доверенности, выдаваемой в порядке, установленном гражданским законодательством Российской Федерации, если иное не предусмотрено настоящим Кодексом. Порядок и правила совершения, а также причины прекращения действия доверенностей устанавливаются положениями раздела 10 Гражданского кодекса Российской Федерации, при этом применение доверенности в отношениях, регулируемых законодательством о налогах и сборах, в рамках Кодекса эти положения не нарушает.

Положениями статьи 188 Гражданского кодекса Российской Федерации установлены основания прекращения действия доверенности. При этом приведенный в указанной статье перечень оснований, по которым доверенность прекращает свое действие, является исчерпывающим и расширительному толкованию не подлежит.

Положениями Кодекса установлены возможность создания доверенности в форме электронного документа, подписанного электронной подписью доверителя (далее – электронная доверенность), – пунктом 3 статьи 29 Кодекса и полномочия ФНС России по утверждению формата и порядка направления электронной доверенности по телекоммуникационным каналам связи (далее – ТКС) – пунктом 5 статьи 80 Кодекса.

На основании вышеуказанных положений Кодекса издан Приказ ФНС России от 30.04.2021 № ЕД-7-26/445@ «Об утверждении формата доверенности, подтверждающей полномочия представителя налогоплательщика (плательщика сбора, плательщика страховых взносов, налогового агента) в отношениях, регулируемых законодательством о налогах и сборах, в электронной форме и порядка ее направления по телекоммуникационным каналам связи».

При этом электронная доверенность применяется как еще один способ создания (совершения) доверенностей, не оказывает влияния на процесс приема от уполномоченных представителей налогоплательщиков налоговой и бухгалтерской отчетности по ТКС, отличного от оказываемого доверенностями, совершенными на бумажных носителях, и не предполагает их отмену.

Внедрение электронной доверенности при осуществлении электронного документооборота с налоговыми органами также не предполагает оснований прекращения действия доверенностей, отличных от установленных в указанной выше статье Гражданского кодекса Российской Федерации.

Таким образом, обязанность, установленная Федеральным законом от 31.07.2023 № 389-ФЗ «О внесении изменений в части первую и вторую Налогового кодекса Российской Федерации, отдельные законодательные акты Российской Федерации и о приостановлении действия абзаца второго пункта 1 статьи 78 части первой Налогового кодекса Российской Федерации» путем внесения изменений в пункт 3 статьи 29 Кодекса, вступающих в силу с 01.03.2024, в соответствии с которыми уполномоченный представитель, указанный в пункте 5.1 статьи 23 Кодекса, осуществляет свои полномочия на основании электронной доверенности, распространяется на доверенности, которые будут совершаться и применяться для подтверждения полномочий уполномоченных представителей после 01.03.2024.

При этом прекращение использования квалифицированных сертификатов сотрудников юридических лиц или замена квалифицированной электронной подписи уполномоченного представителя в рамках действующей доверенности не являются основаниями для прекращения действия доверенности.

Настоящее письмо не является нормативным правовым актом, не влечет изменений правового регулирования налоговых отношений, не содержит норм, влекущих юридические последствия для неопределенного круга лиц, носит информационный характер и не препятствует налогоплательщикам руководствоваться нормами законодательства Российской Федерации о налогах и сборах в понимании, отличающемся от положений настоящего письма.

Figure 1: Original Federal Tax Service letter text (№ ЗГ-3-26/13425 dated 18.10.2023)

Inferred question	Inferred answer	References
Можно ли применять электронные доверенности уполномоченным представителям налогоплательщика-организации?	Да	1. п. 3 ст. 29 НК РФ 2. раздел 10 ГК РФ 3. ст. 188 ГК РФ 4. п. 5 ст. 80 НК РФ 5. п. 5.1 ст. 23 НК РФ 6. Приказ ФНС России от 30.04.2021 № ЕД-7-26/445@

Figure 2: Data inferred from the Federal Tax Service letter № ЗГ-3-26/13425

For the simplicity of evaluation we only consider questions that imply a binary answer (Yes/No). On the one hand, this way of binary classification offers us a very useful evaluation framework. However, it also imposes some significant simplifications on the complexity of a legal discourse. Legal responses, particularly in tax law, often contain a big amount of exceptions, temporal constraints or even contradictions. Therefore, reducing these complex texts to a binary label may cause the losing of semantic context and interpretive richness. In many cases, the final message of the legal authority may only be valid under a very specific set of assumptions, which the binary label does not reflect. Thus, from the overall corpus of over 5000 letters we only select 200 where we are confident enough in the meaningfulness of binary answer.

Legal opinions often involve highly complex reasoning: the introductory part of the text may outline general legal principles, followed by a specification towards the details of the taxpayer's query. And while we aim to associate the binary label with the most appropriate summary of the response, this decision is still interpretive. Therefore, assigning a "Yes" or "No" label may erase the conditional aspect of the opinion.

However, this simplification may favor systems that work well with minimal human supervision. Therefore, we consider our binary classification scheme a necessary abstraction to facilitate evaluation, but it should be modified in the future. Still, we believe that the importance and utility of the dataset we introduce overweighs the drawbacks of the approach.

The basic stats of the dataset are in the Table 1.

Attribute	Value
Number of letters from the Ministry of Finance	131
Number of letters from the Federal Tax Service	78
Answers that can be summarized as "Yes"	89
Answers that can be summarized as "No"	120
Average number of references per letter	4.71
Average number of words per letter	455
Maximum number of words in a letter	1929
Minimum number of words in a letter	89
Median number of words per letter	415
Average number of characters per letter	3580
Maximum number of characters in a letter	15,374
Median number of characters per letter	3255

Table 1. Core statistics of the dataset

2 Prior Datasets Used to Evaluate Reasoning

We have analyzed various papers related to reasoning, including [1-21], and have identified several groups of such datasets. Below is a brief review of such datasets.

2.1 STEM-related datasets

Cobbe et al. [23] introduced GSM8K, a dataset of 8.5K high quality linguistically diverse grade school math word problems. Shi et al. [21] introduced the Multilingual Grade School Math (MGSM) benchmark by manually translating 250 grade-school math problems from the GSM8K dataset into ten typologically diverse languages. Following that, Chen et al. [22] construct the multilingual math reasoning instruction dataset, MGSM8KInstruct, encompassing the same ten languages. Hendrycks et al. [24] introduced MATH, a dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations.

Clark et al. [33] introduced the ARC dataset that contains natural, grade-school science questions authored for human tests. Rein et al. [27] introduced GPQA, a challenging dataset of 448 high-quality and extremely difficult multiple-choice questions written by domain experts in biology, physics, and chemistry. He et al. [28] introduced OlympiadBench, an Olympiad-level bilingual multimodal scientific benchmark, featuring 8,476 problems from Olympiad-level mathematics and physics competitions, including the Chinese college entrance exam.

2.2 Programming datasets

Austin et al. [25] introduced two benchmarks, MBPP and MathQA-Python. The Mostly Basic Programming Problems (MBPP) dataset contains 974 programming tasks, designed to be solvable by entry-level programmers. The MathQA-Python dataset, a Python version of the MathQA benchmark, contains 23914 problems that evaluate the ability of the models to synthesize code from more complex text. Chen et al. [26] introduced HumanEval, an evaluation set intended to measure functional correctness for synthesizing programs from docstrings. Schuster et al. [31] propose a dataset of Python Programming Puzzles (P3). Each puzzle is defined by a short Python program f, and the goal is to find an input that makes f return True. In addition, SWE-bench [65] has recently been proposed as one of the most challenging programming benchmarks. It evaluates models on their ability to solve real-world GitHub issues using information from the associated codebase and documentation.

2.3 First-order logic-based datasets

Saparov and He [43] introduced a synthetic question-answering dataset called PrOntoQA, where each example is generated from a synthetic world model represented in first-order logic.

Tian et al. [41] introduced LogicNLI, an NLI-style dataset that effectively disentangles the target FOL reasoning from commonsense inference and can be used to diagnose LMs from four perspectives: accuracy, robustness, generalization, and interpretability. Yu et al. [36] introduced a Reading Comprehension dataset requiring logical reasoning (ReClor) extracted from standardized graduate admission examinations. Han et al. [34] introduced FOLIO, a human-annotated, logically complex and diverse dataset for reasoning in natural language (NL), equipped with first-order logic (FOL) annotations.

2.4 Synthetic datasets

Yao et al. [29] introduce three tasks requiring non-trivial planning or search: Game of 24, Creative Writing, and 5x5 Mini Crosswords. Game of 24 is a mathematical reasoning challenge, where the goal is to use 4 numbers and basic arithmetic operations (+-*/) to obtain 24. In the creative writing task, the input is 4 random sentences and the output should be a coherent passage with 4 paragraphs that end in the 4 input sentences respectively. Such a task is open-ended and exploratory, and challenges creative thinking as well as high-level planning.

Sinha et al. [35] introduced CLUTRR benchmark that requires that an NLU system infer kinship relations between characters in short stories. Liu et al. [37] introduced LogiQA dataset that is sourced from expert-written questions for testing human Logical reasoning. Liu et al. [38] then amended and re-

annotated it in LogiQA 2.0, increasing the data size, refining the texts with manual translation by professionals, and improving the quality by removing items with distinctive cultural features like Chinese idioms.

Mirzaee and Kordjamshidi [45] introducев SpaRTUN, a synthetic dataset on spatial question answering (SQA) and spatial role labeling (SpRL) to provide a source of supervision with broad coverage of spatial relation types and expressions.

Sakaguchi et al. [47] introduced Winogrande, a large-scale pronoun resolution problem synthetic dataset as a benchmark for commonsense reasoning. Trivedi et al. [44] constructed MuSiQue-Ans, a synthetic multihop QA dataset with 25K 2–4 hop questions by systematically combining pairs of single-hop questions that are connected.

2.5 Datasets in Russian

Fenogenova et al. introduced two Russian machine reading comprehension (MRC) datasets, called MuSeRC and RuCoS, which require reasoning over multiple sentences and commonsense knowledge to infer the answer. The former follows the design of MultiRC [55], while the latter is a counterpart of the ReCoRD dataset [56]. Shavrina et al. [52] introduced a Russian general language understanding evaluation benchmark – Russian SuperGLUE, organized similarly to SuperGLUE [54] and including MuSeRC and RuCoS. Taktasheva et al. [46] proposed TAPE (Text Attack and Perturbation Evaluation), a benchmark that includes six more complex NLU tasks for Russian, covering multi-hop reasoning, ethical concepts, logic and commonsense knowledge.

2.6 Other datasets

Srivastava et al. [42] introduced BIG-bench that consists of 204 tasks, contributed by 450 authors across 132 institutions. Task topics are diverse, drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development etc. Suzgun et al. [30] select a suite of 23 challenging BIG-Bench tasks for which prior language model evaluations did not outperform the average human-rater and call it BIG-Bench Hard (BBH).

Another prominent benchmark is ARC AGI [66]. It emphasizes general intelligence and abstract reasoning skills through a set of a diverse tasks that for now remain unsolved by many modern systems.

Thorne et al. [51] introduced a dataset for verification against textual sources, FEVER: Fact Extraction and VERification. It consists of claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. Yang et al. [50] introduced HotpotQA, a dataset with Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) sentence-level supporting facts required for reasoning are provided, allowing QA systems to reason with strong supervision and explain the predictions; (4) factoid comparison questions are provided to test QA systems' ability to extract relevant facts and perform necessary comparison.

Bisk et al. [49] introduced the task of physical commonsense reasoning and a corresponding benchmark dataset Physical Interaction: Question Answering or PIQA. Cui et al. [48] introduced MuTual, a dataset for Multi-Turn dialogue Reasoning based on Chinese student English listening comprehension exams.

3 Data Processing Pipeline

In this section, we describe the steps performed to build the dataset. In short, the steps are:

- Downloading
- Candidate selection
- Answer binarity validation
- Question extraction
- Postprocessing
- Extracting legal sources

3.1 Downloading

Any document types mentioned below were downloaded from the locations detailed in the Table 2.

3.2 Candidate Selection and Answer Binarity Validation

Any Ministry of Finance letters from departments other than the tax policy department were removed from consideration. The remaining letters were run through GPT-4o mini with a prompt that determines whether the answer can be summarized as "yes" or "no" (see Appendix A).

3.3 Question Extraction and Manual Postprocessing

The letters that were determined by the prompt as implying a binary answer were run through GPT-4o mini with a prompt that created the questions implied by the answer given by the respective authority (see Appendix A). Since the ability to summarize a response as "Yes" or "No" largely depends on the phrasing of the question, two distinct questions were generated for each letter. One of them was generated in such a way that the answer could be expressed as a clear "Yes", and the other in such a way that the answer could be expressed only as "No". Then the letters and questions have been checked manually by the team members. During the review process, team members selected the most appropriate of the two generated questions, making adjustments if necessary. Where necessary, the question was completely rewritten. We deliberately allowed the balance to be shifted towards the questions that assume the answer "No", since according to our observations, refuting the assumption given in the question requires more bold reasoning from a LLM. It is also worth noting that some of the compiled question-answer pairs were then manually removed from the dataset, since their content was limited to links to some regulatory and legal acts of the Russian Federation, and therefore they were of little value to the dataset.

Source type	Download location
Письма Минфина, Письма ФНС, Приказы ФНС	https://www.audar-info.ru
Гражданский кодекс, Бюджетный кодекс, Трудовой кодекс	http://pravo.gov.ru
Налоговый кодекс	https://nalog.garant.ru/fns/nk
Постановления правительства, Распоряжения правительства, Указы президента, Распоряжения президента, Федеральные законы, Федеральные конституционные законы	http://government.ru/docs/all
Разъяснения с сайта buh.ru	https://buh.ru/articles
Разъяснения с сайта 1С	https://its.1c.ru/db/newscomm
Статьи kontur-extern.ru	https://www.kontur-extern.ru/info/

Table 2: Legal data sources

3.4 Extracting legal references

To benchmark the search quality in RAG etc. one needs to have a gold standard on mentions of legal sources in the answer (legal references). To create one, we automatically extracted any legal references with a script and then selectively checked the completeness of references extracted. Whenever the manual testing revealed incomplete list of references extracted, the script was modified and the cycle was repeated.

4 Baseline systems

4.1 LLM and RAG

We used Mixtral 8x7B [57], Llama 3.3 70B [58] and GPT-4o mini [62] as baseline LLMs. We have also built a baseline Retrieval-Augmented Generation (RAG) system [63] that combines vector search using E5 [59] dense vector embeddings with generation using the above LLMs.

4.2 Finetuning E5

To improve the search quality, we used contrastive pre-training of E5 [59]. We pre-train the model on positive examples in the form of pairs (user query - relevant documents), and on negative examples in the form of pairs (user query - random irrelevant documents). This is the same approach as for the pre-training stage of E5 [59]. We use InfoNCE [60], [61] as a contrastive loss function, and use in-batch negatives, i.e. when answers from all other pairs of the batch are taken as negative examples for a question. The quality of such training usually increases with the batch size.

We have created two types of pairs: question – text from legal reference, and heading – text. Based for the first one we have created the following types of pairs:

- Question title - text from reference
- Question text - text from reference
- Compressed question text - text from reference
- Question title - compressed text from reference
- Question text - compressed text from reference
- Compressed question text - compressed text from reference

Long texts exceeding the model context were summarized using IlyaGusev/rut5_base_sum_gazeta.

Heading – text included pairs of heading - article text, or paragraph text - article subparagraph text, if such a hierarchy exists in the document, the title of the law / regulation and its text from the following sources:

- Tax Code of the Russian Federation
- Civil Code of the Russian Federation
- Labor Code of the Russian Federation
- Letters from the Ministry of Finance
- Letters from the Federal Tax Service
- Federal laws
- Government decrees

This data set was split into training (23501) and validation (1201) samples randomly (ratio 0.95 / 0.05). Finetuning was performed on NVIDIA RTX 4090 GPU. We used the hyperparameters recommended for fine-tuning in the E5 [59]. Despite the mixed precision, the batch size more than 5 overflowed VRAM. To artificially increase the batch size, gradient accumulation was used.

Validation consists of:

- looking at InfoNCE loss on validation samples of task 1 (question titles / full texts of questions separately) and task 2;
- using the updated version of the model, we collect a database (faiss) from the link texts of the training sample of task 1, and perform a semantic search on questions from the validation sample of task 1:
  - looking at Recall@10, Recall@20, Recall@40;
  - for all types of documents together and for each separately - Tax Code of the Russian Federation, Civil Code of the Russian Federation, Letters of the Ministry of Finance, Letters of the Federal Tax Service, Orders of the Federal Tax Service, Federal Laws, Government Resolutions (except for the Labor Code of the Russian Federation and Courts).

Figures 3-5 demonstrate the finetuning process in the terms detailed above.

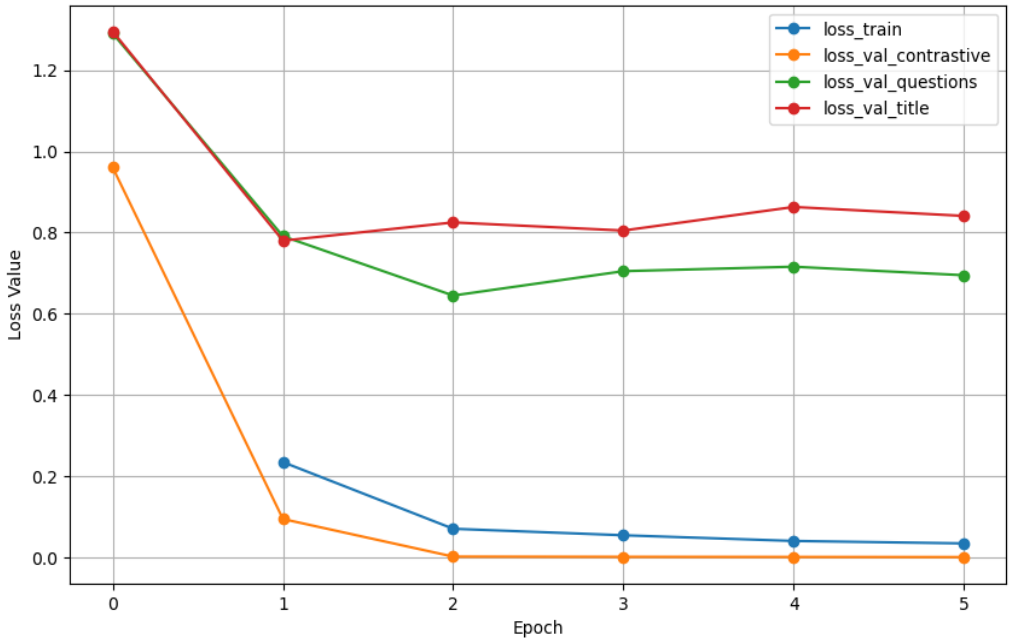

Figure 3: Losses for E5 finetuning

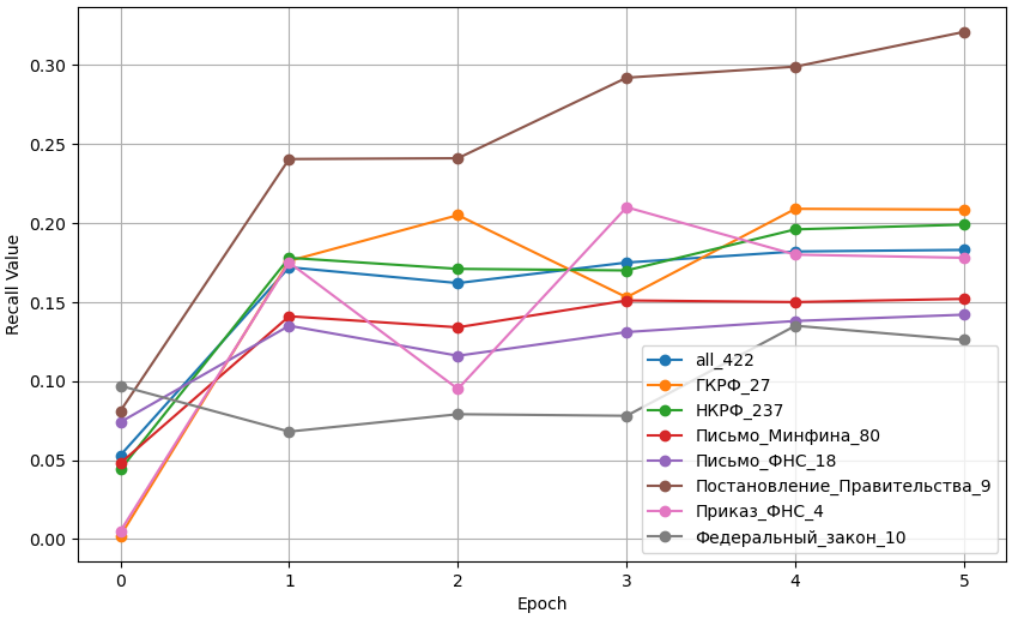

Figure 4: Recall @10 for the validation set of E5 finetuning

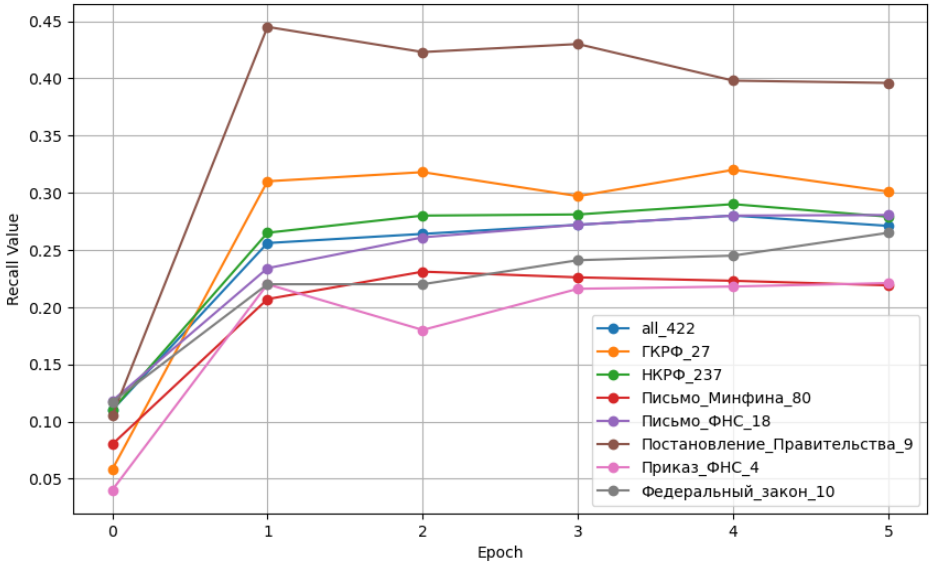

Figure 5: Recall @40 for the validation set of E5 finetuning

5 Evaluation results

We have evaluated four experimental setups:

- **Simple**: Questions from the dataset are presented to the model in their original form, accompanied by a straightforward prompt within a Chain-of-Thought framework [64]. A positive motivational tone is included to guide responses. Prompts are presented in Appendix A.
- **"Perfect" RAG**: Questions are supplemented with verbatim excerpts from the documents referenced in the corresponding Letters, ensuring that the model has access to the same information available to legal experts that wrote those Letters.
- **Base RAG:** This setup employs the E5-large model with a preconstructed vector database of relevant documents. The top 15 most relevant text chunks, retrieved based on the question, are provided to the model as supporting context alongside with the question.
- **Fine-tuned RAG:** Similar to the Base RAG setup, but the retrieval step utilizes our fine-tuned E5 model for improved relevance in retrieving supporting documents.

The accuracy scores for each setup are summarized in Table 3.

Model	Simple	"Perfect" RAG	Base RAG	Fine-tuned RAG
Mixtral-8x7B-Instruct-v0.1	0.566	0.684	0.658	0.606
Llama-3.3-70B-Instruct	0.625	0.775	0.766	0.746
GPT-4o mini	0.644	0.748	0.770	0.760

Table 3: Accuracy scores on RuTaR dataset in various experimental setups

These results highlight the consistent performance improvement achieved through the inclusion of RAG setups, while also pointing to potential areas for improvement in future, more advanced models. Interestingly, while the fine-tuned RAG setup demonstrates a significant enhancement in retrieval quality—evidenced by the substantially higher recall metrics shown in Figure 6 (e.g., recall@15 of 72% for fine-tuned versus 49% for Base RAG)—its accuracy is paradoxically slightly lower than that of the Base RAG setup. Even more intriguing is that the Perfect RAG does not help GPT-4o mini to improve over the Base RAG setup. This discrepancy indicates that retrieval quality alone does not fully account for downstream performance. Instead, factors such as the relevance ranking of retrieved chunks, the diversity of retrieved information, or the model's ability to integrate and reason over retrieved content may have played a more significant role.

Further research is needed to better understand this phenomenon. Potential avenues for analysis include a qualitative review of the retrieved chunks' characteristics and an examination of how these influence the reasoning and decision-making process of the model. Such insights would provide valuable guidance for enhancing fine-tuning approaches and optimizing the end-to-end RAG pipeline.

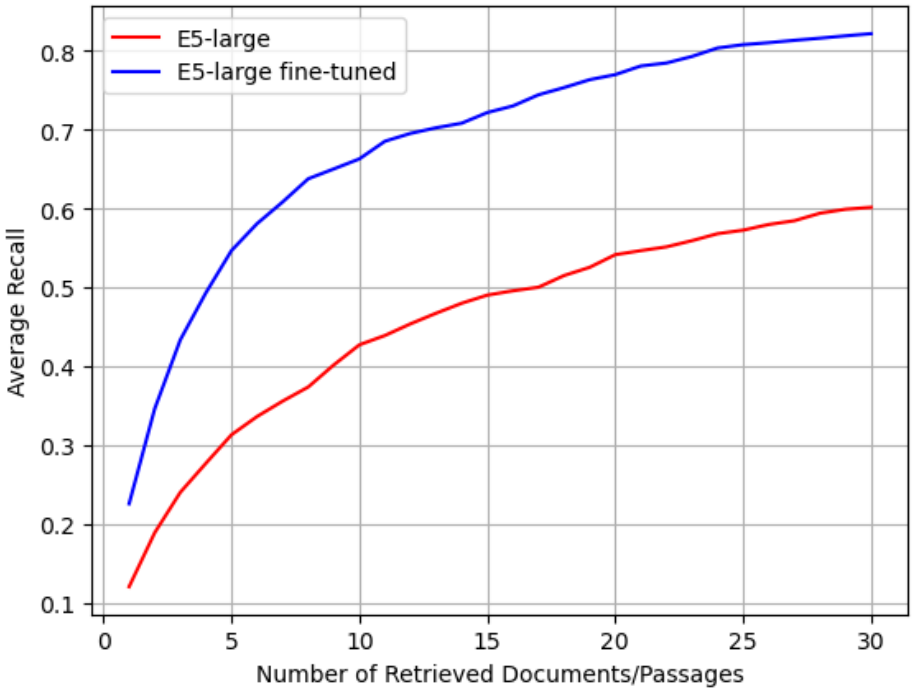

Figure 6: Average Recall of Retrieval Models on Questions from the RuTaR Dataset

6 Discussion

The presented dataset consists of questions that are hard for human professionals. That's why they come to government bodies for authoritative answers. The dataset is equally complicated for LLMs and vector search. Current popular LLMs built without specific attention to reasoning exhibit accuracy of 57% to 65% out-of-the-box, and 66% to 77% with the best RAG setup. Given the apparent simplicity of the setup that implies the accuracy of random choice of 50%, the dataset allows for validating significant improvement in both reasoning and retrieval capabilities of LLMs.

Across the evaluated systems, the most frequent error pattern was the systems failure to deliver a final binary verdict, despite partially correct or relevant reasoning. A likely explanation for that may be that, while the models demonstrate basic legal understanding, they often don't complete their reasoning processes when lacking the needed information from the retrieved documents. In cases where a RAG pipeline was employed (both base and ideal), some models still failed to reach a conclusion. This may be attributed to either retrieval noise (irrelevant or tangentially related documents), or the models' difficulty in synthesizing multiple legal references into a coherent output. Conversely, fine-tuned RAG pipelines significantly reduced such errors, suggesting that both retrieval quality and task-specific tuning play a critical role in guiding models toward conclusive legal reasoning.

Acknowledgements

The authors are grateful to their colleagues from NTR Labs computational linguistics and RL divisions for the support and discussions. This work is also largely inspired by discussions with our colleagues from Norilsky Nikel innovation and IT divisions including Alexey Testin, Anton Burkov, Kirill Maidanik, Sergey Greben and Gleb Volk.

References

[1] Janice Ahn, Rishu Verma, et al. Large Language Models for Mathematical Reasoning: Progresses and Challenges // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, P. 225–237, St. Julian's, Malta. Association for Computational Linguistics.

[2] Sohee Yang, Elena Gribovskaya, et al. Do Large Language Models Latently Perform Multi-Hop Reasoning? // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), P. 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

[3] Jie Huang, Xinyun Chen, et al. Large Language Models Cannot Self-Correct Reasoning Yet // The Twelfth International Conference on Learning Representations, 2024, https://openreview.net/forum?id=IkmD3fKBPQ

[4] Aviral Kumar, Vincent Zhuang, et al. Training Language Models to Self-Correct via Reinforcement Learning // Computing Research Repository. — 2024. — Vol. arXiv:2409.12917

[5] Pranav Putta, Edmund Mills, et al. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents // Computing Research Repository. — 2024. — Vol. arXiv:2408.07199 https://arxiv.org/abs/2408.07199

[6] Aditya Kalyanpur, Kailash Karthik Saravanakumar, et al. LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic// Computing Research Repository. — 2024. — Vol. arXiv:2406.17663 https://arxiv.org/abs/2406.17663

[7] Chaojie Wang, Yanchen Deng , et al. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning Computing Research Repository. — 2024. — Vol. arXiv:2406.14283 https://arxiv.org/abs/2406.14283

[8] Ye Tian, Baolin Peng, et al. Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing Computing Research Repository. — 2024. — Vol. arXiv:2404.12253 https://arxiv.org/abs/2404.12253

[9] Leonardo Ranaldi, Giulia Pucci, et al.. 2024. Empowering Multi-step Reasoning across Languages via Program-Aided Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12171–12187, Miami, Florida, USA. Association for Computational Linguistics.

[10] Charlie Snell, Jaehoon Lee, , et al. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. Computing Research Repository. — 2024. — Vol. arXiv:2408.03314.

[11] Yoshua Bengio, From System 1 Deep Learning to System 2 Deep Learning https://youtu.be/T3sxeTgT4qc

[12] Open AI Team. Learning to Reason with LLMs. Technical Report'24. September 12, 2024 https://openai.com/index/learning-to-reason-with-llms/

[13] Yu Zhao, Huifeng Yin , et al. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions, Computing Research Repository. — 2024. — Vol. arXiv:2411.14405

[14] Zhen Huang, Haoyang Zou, et al. O1 Replication Journey -- Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson Computing Research Repository. — 2024. — Vol. arXiv:2411.16489 https://arxiv.org/abs/2411.16489

[15] Yiwei Qin, Xuefeng Li, , et al. O1 replication journey: A strategic progress report–part 1. Computing Research Repository. — 2024. — Vol. arXiv:2410.18982.

[16] LLaMaO1 Team. 2024. Llamao1. Github https://github.com/SimpleBerry/LLaMA-O1

[17] OpenO1 Team. 2024. Openo1. Github https://github.com/Open-Source-O1/Open-O1

[18] kimi. 2024. k0math. website https://kimi.moonshot.cn/

[19] kunlun. 2024. skyworko1. website https://ai-bot.cn/skywork-o1/

[20] Yuhao Dong, Zuyan Liu, et al. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models Computing Research Repository. — 2024. — Vol. arXiv:2411.14432 https://arxiv.org/abs/2411.14432

[21] Freda Shi, Mirac Suzgun, et al. Language models are multilingual chain-of-thought reasoners The Eleventh International Conference on Learning Representations 2023, https://openreview.net/forum?id=fR3wGCk-IXp

[22] Nuo Chen, Zinan Zheng, et al. Breaking Language Barriers in Multilingual Mathematical Reasoning: Insights and Observations. // Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

[23] Karl Cobbe, Vineet Kosaraju, et al. Training Verifiers to Solve Math Word Problems Computing Research Repository. — 2021. — Vol. arXiv:2110.14168 https://arxiv.org/abs/2110.14168

[24] Dan Hendrycks, Collin Burns, et al. Measuring Mathematical Problem Solving With the {MATH} Dataset // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021, https://openreview.net/forum?id=7Bywt2mQsCe

[25] Jacob Austin, Augustus Odena, et al. Program Synthesis with Large Language Models 2021, Computing Research Repository. — 2021. — Vol. arXiv:2108.07732, https://arxiv.org/abs/2108.07732

[26] Mark Chen, Jerry Tworek, et al. Evaluating Large Language Models Trained on Code Computing Research Repository. — 2021. — Vol. arXiv: 2107.03374 https://arxiv.org/abs/2107.03374

[27] David Rein, Betty Li Hou, et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark Computing Research Repository. — 2023. — Vol. arXiv:2311.12022 https://arxiv.org/abs/2311.12022
[28] Chaoqun He, Renjie Luo, et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems Computing Research Repository. — 2024. — Vol. arXiv:2402.14008, https://arxiv.org/abs/2402.14008
[29] S. Yao, D. Yu, et al. Tree of thoughts: Deliberate problem solving with large language modelsm // Advances in Neural Information Processing Systems, vol. 36, 2024.
[30] M. Suzgun, N. Scales, et al. Challenging big-bench tasks and whether chain-of-thought can solve them // Findings of the Association for Computational Linguistics: ACL 2023, P. 13003–13051, 2023.
[31] T. Schuster, A. Kalyan, et al. Programming puzzles // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.
[32] A. T. K. Patrick Haluptzok, Matthew Bowers. Language models can teach themselves to program better // Eleventh International Conference on Learning Representations (ICLR), 2023.
[33] Peter Clark, Isaac Cowhey, et al. Think you have solved question answering? Try ARC, the AI^2 reasoning challenge. Computing Research Repository. — 2018. — Vol. arXiv:1803.05457.
[34] Simeng Han, Hailey Schoelkopf, et al. FOLIO: Natural Language Reasoning with First-Order Logic. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, P. 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.
[35] Koustuv Sinha, Shagun Sodhani, et al. CLUTRR: A diagnostic benchmark for inductive reasoning from text. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), P. 4506–4515, Hong Kong, China. Association for Computational Linguistics
[36] Weihao Yu, Zihang Jiang, et al. ReClor: A reading comprehension dataset requiring logical reasoning. // International Conference on Learning Representations.
[37] Jian Liu, Leyang Cui, et al. LogiQA: a challenge dataset for machine reading comprehension with logical reasoning // Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, P. 3622–3628
[38] Hanmeng Liu, Jian Liu, et al. LogiQA 2.0 — an improved dataset for logical reasoning in natural language understanding // IEEE/ACM Transactions on Audio, Speech, and Language Processing, P. 1–16, 2023. doi: 10.1109/TASLP.2023.3293046.
[39] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. // Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, P. 3882–3890.
[40] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, P. 3621–3634, Online. Association for Computational Linguistics.
[41] Jidong Tian, Yitian Li, et al. Diagnosing the firstorder logical reasoning ability through LogicNLI. // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, P. 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
[42] Aarohi Srivastava, Abhinav Rastogi, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Computing Research Repository. — 2024. — Vol. arXiv:2206.04615
[43] Abulhair Saparov and He He. Language Models Are Greedy Reasoners: A systematic formal analysis of chain-of-thought. // International Conference on Learning Representations 2023
[44] Harsh Trivedi, Niranjan Balasubramanian, et al. ♫ MuSiQue: Multihop Questions via Single-hop Question Composition. // Transactions of the Association for Computational Linguistics, 10:539–554.
[45] Roshanak Mirzaee and Parisa Kordjamshidi. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, P. 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
[46] Ekaterina Taktasheva, Alena Fenogenova, et al. TAPE: Assessing Few-shot Russian Language Understanding. // Findings of the Association for Computational Linguistics: EMNLP 2022, P. 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
[47] Keisuke Sakaguchi, Ronan Le Bras, et al. Winogrande: An adversarial Winograd schema challenge at scale. Computing Research Repository. — 2019. — Vol. arXiv:1907.10641.
[48] Leyang Cui, Yu Wu, et al. MuTual: A dataset for multi-turn dialogue reasoning. // Proceedings of the 58th Conference of the Association for Computational Linguistics. Association for Computational Linguistics, 2020.
[49] Yonatan Bisk, Rowan Zellers, et al. PIQA: Reasoning about physical commonsense in natural language. // Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[50] Zhilin Yang, Peng Qi, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, P. 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
[51] James Thorne, Andreas Vlachos, et al. FEVER: a Large-scale Dataset for Fact Extraction and VERification. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), P. 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
[52] Tatiana Shavrina, Alena Fenogenova, et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), P. 4717–4726, Online. Association for Computational Linguistics.
[53] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. // Proceedings of the 28th International Conference on Computational Linguistics, pages 6481–6497, Barcelona, Spain (Online). International Committee on Computational Linguistics.
[54] Alex Wang, Yada Pruksachatkun, et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 294, 3266–3280.
[55] Daniel Khashabi, Snigdha Chaturvedi, et al. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), P. 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
[56] Zhang, Sheng, Xiaodong Liu, et al. Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885 (2018).
[57] Albert Q. Jiang, Alexandre Sablayrolles, et al. Mixtral of experts. https://arxiv.org/abs/2401.04088
[58] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/
[59] Wang, Liang, Nan Yang, et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training. Computing Research Repository. — 2022. — Vol. arXiv:/2212.03533
[60] Ting Chen, Simon Kornblith, et al. A simple framework for contrastive learning of visual representations. // Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 149, 1597–1607.
[61] Oord, Aäron van den, Yazhe Li and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. Computing Research Repository. — 2018. — Vol. arXiv:1807.03748
[62] Open AI Team. GPT-4o mini: advancing cost-efficient intelligence. Technical Report'24. July 18, 2024 https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[63] Patrick Lewis, Ethan Perez, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. // Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793, P. 9459–9474.
[64] Wei, Jason, Xuezhi Wang, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Computing Research Repository. — 2022. — Vol. arXiv:2201.11903. https://arxiv.org/pdf/2201.11903.
[65] Jimenez, Carlos E., Yang, John, Wettig, Alexander, Yao, Shunyu, Pei, Kexin, Press, Ofir, Narasimhan, Karthik. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? Computing Research Repository. — 2024. — Vol. arXiv:2310.06770. URL: https://arxiv.org/abs/2310.06770
[66] Chollet, François. On the Measure of Intelligence. Computing Research Repository. — 2019. — Vol. arXiv:1911.01547. URL: https://arxiv.org/abs/1911.01547

Appendix A. Prompts used.

Prompt that determines whether the answer can be summarized as "yes" or "no"

```
system_prompt = (
    "Вы - юридический помощник, которому нужно определить, можно
ли суммаризировать представленный ответ на вопрос как 'да' или
'нет'. "
    "Если это возможно, ответьте 'Да' или 'Нет', в зависимости от
смысла. "
    "Если суммаризация невозможна, ответьте 'Недвоичный ответ'."
    "Возможные варианты вашего ответа: 'Да','Нет','Недвоичный от-
вет'. Пожалуйста, больше ничего не пишите.”

user_prompt = f"Вопрос: {question}\nОтвет: {answer}\nМожно ли
суммаризировать ответ как 'Да' или ‘Нет'?"
```

The letters that were determined by the prompt as implying a binary answer were run through GPT-4o with a prompt that created the questions implied by the answer given by the respective authority (see Appendix B).

```
Вариант 1.

system_prompt = (
    "Вы — квалифицированный юридический помощник, который генери-
рует бинарные вопросы (такие, на которые можно ответить 'да' или
'нет') на основе предоставленных ответов. "
    "Сгенерированные вопросы должны максимально напоминать во-
просы, задаваемые реальными пользователями и строго соответство-
вать содержанию ответа."
    "Каждый вопрос должен быть составлен таким образом, чтобы от-
вет на него можно было суммаризировать как 'да' или 'нет'."
    "Вопросы должны быть вежливыми, корректными и при необходимо-
сти включать в себя релевантную юридическую терминологию."
    "Вы не должны писать ничего, кроме текста самого вопроса. В
том числе, пожалуйста, избегайте приветствий, прощаний, вваодных
конструкций и комментариев."
)

user_prompt = f"Пожалуйста, сгенерируйте бинарный вопрос на ос-
нове следующего ответа:\n{answer}"
```

Вариант 2.

```
system_prompt = (
    "Вы — квалифицированный юридический помощник, который генерирует бинарные вопросы (такие, на которые можно ответить 'да' или 'нет') на основе предоставленных ответов. "
    "Сгенерированные вопросы должны максимально напоминать вопросы, задаваемые реальными пользователями и строго соответствовать содержанию ответа."
    "Каждый вопрос должен быть составлен таким образом, чтобы ответ на него можно было суммаризировать только как 'нет'."
    "Вопросы должны быть вежливыми, корректными и при необходимости включать в себя релевантную юридическую терминологию."
    "Вы не должны писать ничего, кроме текста самого вопроса. В том числе, пожалуйста, избегайте приветствий, прощаний, вводных конструкций и комментариев."
)

user_prompt = f"Пожалуйста, сгенерируйте бинарный вопрос на основе следующего ответа:\n{answer}. Сгенерируйте вопрос таким образом, чтобы ответ на него был отрицательным.”
```