Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09463-9

ORIGINAL RESEARCH

Enhancing legal question answering with data generation and knowledge distillation from large language models

Paolo Italiani[1] · **Gianluca Moro**[1] · **Luca Ragazzi**[1]

Accepted: 24 April 2025

Abstract

Legal question answering (LQA) relies on supervised methods to automatically handle law-related queries. These solutions require a substantial amount of carefully annotated data for training, which makes the process very costly. Although large language models (LLMs) show promise in zero-shot QA, their computational demands limit their practical use, making specialized small language models (SLMs) more favorable. Furthermore, the growing interest in synthetic data generation has recently surged, spurred by the impressive generation capabilities of LLMs. This paper presents ACE-ATTORNEY, an LLM distillation approach devised to develop LQA data and supervised models without human annotation. Given a textual prompt, a frozen LLM generates artificial examples that are used as knowledge to train a student SLM with an order of magnitude fewer parameters. Taking into account a realistic retrieval-based scenario to fetch the correct document for answer generation, we propose Selective Generative Paradigm, a novel approach designed to improve retrieval efficacy. Extensive experiments demonstrate the effectiveness and efficiency of distilled models on SYN- LEQA, our human-free synthetic dataset, and a public expert-annotated corpus. Notably, by using only a few dozen training samples, our best SLM achieves LLM-comparable performance with ≈1200% less CO_2 emissions. The data and the code to fully reproduce our results are available at https://github.com/disi-unibo-nlp/ace-attorney.

Keywords Large language models · Legal question answering · Knowledge distillation · Synthetic data generation

All authors contributed equally to this paper.

✉ Gianluca Moro
gianluca.moro@unibo.it

Paolo Italiani
paolo.italiani@unibo.it

Luca Ragazzi
l.ragazzi@unibo.it

[1] Department of Computer Science and Engineering - DISI, University of Bologna, Via dell'Universit'a 50, I-47522 Cesena, Italy

Published online: 16 July 2025

1 Introduction

The rapid introduction and growing availability of novel digitized legal texts result in an overload of information for legal professionals. Acting as an expert assistant, legal question answering (LQA) could help quickly capture the main points of new cases (Martinez-Gil 2023), similar to summarization systems (Ragazzi et al. 2024). However, the need for accurate training data and its associated annotation cost represent a barrier for emerging applications. Consequently, there is a pressing demand for LQA data and models to streamline the deployment of production-ready solutions in high-value domains such as law (Louis et al. 2023).

Automated data generation has recently attracted enormous interest due to the remarkable generative capacity of large language models (LLMs). On the downside, the huge size of such models poses notable deployment obstacles (e.g., >190 GB of GPU RAM is needed to serve a 180B LLM with 8-bit quantization Almazrouei et al. 2023). Consequently, given such prohibitive costs, knowledge distillation (KD) (Hinton et al. 2015) has emerged as a strategy, creating task-specific small language models (SLMs) trained on synthetic data produced by LLMs (Meng et al. 2022; Ye et al. 2022; Gao et al. 2023). This paradigm enables the creation of cost-efficient student models, affordable for most product teams, which often outperform prompt-based zero-shot LLMs despite having order-of-magnitude fewer parameters (Zhou et al. 2023). However, previous work on KD has focused mainly on text classification, neglecting generative tasks such as QA. Furthermore, upon closer inspection, existing LQA datasets[1] exhibit at least one of the following limitations: (i) they concern specialized domains with constrained scope, such as tax law (Holzenberger et al. 2020); (ii) they are restricted to multiple choice (Zheng et al. 2021; Bongard et al. 2022) or few-word replies (Sovrano et al. 2021); (iii) the answer is a verbatim extraction from the context (Ravichander et al. 2019; Ahmad et al. 2020).

Motivated by the power of LLM text generation and the scarcity of LQA contributions, we present ACE- ATTORNEY (Fig. 1), an LLM KD approach to automate the generation of LQA datasets and models. First, we use a frozen instruction-tuned LLM to create a comprehensive synthetic QA dataset via carefully designed prompts on legal corpora, filling the dearth of publicly available LQA datasets.[2] This phase is enhanced by a data refinement process, where specific instances are modified to cover realistic scenarios in which an incorrect context is used to answer a given question. Subsequently, a student SLM is trained on the artificial question–answer pairs, operating in a pure zero-shot setting without involving human annotations throughout the entire pipeline. We based our analysis on a realistic scenario that involves retrieving the correct document within a cluster to answer a determined question. To this end, taking inspiration from existing works on relevance assessment (Faggioli et al. 2023; MacAvaney and Soldaini 2023; Thomas et al. 2024), we present Selective Generative Paradigm (SGP), a methodology grounded in iterative refinement (Domeniconi et al. 2014, 2015), specifically designed to improve the effectiveness of the retrieval

[1] As widely accepted by the community to track advance progress, we analyzed contributions for English.

[2] Note that our study aims to generate data ex novo, going beyond the purpose of data augmentation (Ghosh et al. 2023).

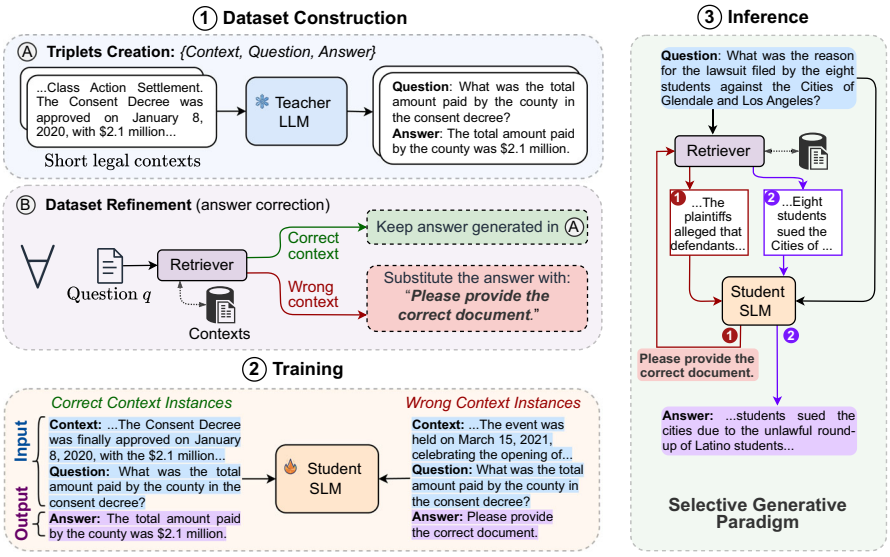


Fig. 1 ACE- ATTORNEY overview. (1) A frozen LLM generates {context, question, answer} triplets from short legal contexts. For each question, we retrieve the corresponding context, manually adjusting the answer if the context is incorrect. (2) We then train an SLM on these triplets. (3) During inference, we employ a Selective Generative Paradigm, where the SLM provides an answer based on the relevance of the retrieved document

module. Thanks to our prior dataset refinement process, the generative model—tasked with answering the question—can directly evaluate the pertinence of the retrieved document and request a new one if necessary. ACE- ATTORNEY accommodates models of different sizes without presumption on their architecture, fulfilling the available hardware infrastructure.

For experimentation, we consider privacy policies and civil rights lawsuits as the raw legal knowledge to build SYN- LEQA, a new synthetic LQA dataset originated by LLM KD. We then fine-tune multiple million-scale SLMs on it and perform in- and cross-domain experiments on a real human-annotated LQA dataset.

In summary, our main contributions are the following:

- We explore synthetic data generation for legal applications, presenting ACE- ATTORNEY, a model-agnostic approach that operates via LLM KD to automate the production of data and small specialized models. With our approach, we create SYN- LEQA, the first high-quality synthetic dataset for generative LQA.
- We propose SGP, which allows the generator and the retriever to work jointly to select the correct document to answer the question.
- By benchmarking multiple neo-released billion-scale LLMs in LQA, we show that our distilled SLMs—despite having order-of-magnitude fewer parameters—outperform LLMs with $\approx$1200% less CO_2 emissions, even obtaining LLM-comparable results in a real-case scenario of just a few dozen training samples.

2 Related work

Legal Question Answering LQA has been a long-standing challenge for legal intelligence applications (Kim et al. 2016), fueled by the overwhelming influx of new legal information. Consequently, to formulate and facilitate practical remedies, several datasets have been proposed and publicly released. CASEHOLD (Zheng et al. 2021) collects more than 53K multiple-choice questions to identify the relevant holding of a cited case. PRIVACYQA (Ravichander et al. 2019) aggregates 1750 questions on privacy policies of mobile applications. POLICYQA (Ahmad et al. 2020) provides 714 human-annotated inquiries for a wide range of privacy practices. To the best of our knowledge, there are no open-access datasets that offer generative answers, although this is often a user requirement for lawyers (Zhang et al. 2023).

Synthetic Data Generation Several works focused on the fabrication of artificial data using generative models. Early attempts take advantage of human annotations to generate data samples and train models in a semi-supervised fashion (Anaby-Tavor et al. 2020; Kumar et al. 2020; Puri et al. 2020; Lee et al. 2021). As LLMs have become increasingly popular, they have also gained attention for generating task-specific examples (Yoo et al. 2021; Bonifacio et al. 2022; Carranza et al. 2023; Guo et al. 2023). However, these solutions are still based on expert-made data. We advocate for a recent research pathway that explores a pure zero-shot scenario, fueled by LLM KD to fine-tune cost-efficient SLMs. Schick and Schütze (2021) use LLMs to generate labeled text pairs for the semantic textual similarity task. Meng et al. (2022) produce samples using generation probability and regularization techniques for text classification. Yu et al. (2023) use complex attributed prompts, creating diverse and attributed artificial data. Ye et al. (2022) push this concept to its extreme by training LSTMs on synthetic data. However, as far as we know, data generation has never been explored for legal NLP tasks, despite the benefit of bypassing the excessive annotation effort required by in-domain experts to create high-quality labeled examples.

3 ACE-ATTORNEY

We introduce ACE- ATTORNEY, a new approach devised to develop LQA-specific data and supervised models through LLM KD.

Task Definition LQA can be formally defined as follows. Let $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ be a corpus of legal documents. Given a single-sentence question $\mathbf{q}$, a retriever $\mathcal{R}$ is asked to sort the documents in $\mathbf{D}$ by relevance: $\{\mathbf{d}'_1, \ldots, \mathbf{d}'_n\} = \mathcal{R}(\mathbf{D} \mid \mathbf{q})$. Then, a generative model $\mathcal{G}$ produces the answer $\mathbf{a}$ using both $\mathbf{q}$ and $\mathbf{d}'_1$, namely $\mathbf{a} = \mathcal{G}(\mathbf{q}, \mathbf{d}'_1)$.

3.1 SYN-LEQA

Motivated by the inaccessibility of publicly available LQA datasets, this work focuses on the generation of examples $\mathcal{E} = <\mathbf{q}, \mathbf{d}'_1, \mathbf{a}> \mid \mathbf{d}'_1 \in \mathbf{D}$, composing SYN- LEQA,

Table 1 Dataset statistics for both SYN- LEQA and PRIVACYQA$_{gen}$

Set	# Samples	# Question Words	# Answer Words	# Doc Words
SYN- LEQA				
Train	3854	27.55	34.50	115.90
Val	481	27.66	34.46	118.39
Test	483	27.82	34.32	116.31
PRIVACYQA$_{gen}$				
Test	431	9.37	42.48	116.70

The word count is computed with the NLTK tokenizer (Bird et al. 2017)

our synthetic LQA dataset for training $\mathcal{G}$. Table 1 shows the statistics of our dataset, while Table 2 displays two random examples.

Challenges This data generation process faces several challenges. C1 The documents in $\mathbf{D}$ leveraged to generate examples $\mathcal{E}_1, \ldots, \mathcal{E}_{|\mathcal{E}|}$ should not be overly similar to ensure generalizability across unseen data and domains, but should also not be too dissimilar to sidestep succumbing to a trivial retrieval task. C2 The questions must provide sufficient information for $\mathcal{R}$ to accurately retrieve the corresponding document in $\mathbf{D}$. C3 The answers should be coherent and factual w.r.t. the question and the source document, respectively.

Source Data For the fabrication of SYN- LEQA, we identify two diverse legal sources in $\mathbf{D}$, so that we can tackle C1. **(1)** *Civil rights lawsuits*: we consider short summaries of MULTI- LEXSUM (Shen et al. 2022), a legal corpus for multi-document summarization, consisting of a description of the background, parties involved, and outcome of the case. **(2)** *Privacy policies*: we aggregate the policies provided by the training set of PRIVACYQA (Ravichander et al. 2019) and POLICYQA (Ahmad et al. 2020), which are contractual agreements that bind companies to their clients. To align with the input length of short summaries in MULTI- LEXSUM and increase the number of samples, we divided the original policies into chunks of a maximum of four sentences.

LLM Annotation We use LLAMA- 2- 13B as LLM teacher $\mathcal{M}$.[3] Specifically, given a document $\mathbf{d}_i$, we provide $\mathcal{M}$ with a curated prompt template $\mathcal{T}(\cdot)$ to generate the question–answer pairs, formally defined as $(\mathbf{q}, \mathbf{a}) = \mathcal{M}(\mathcal{T}(\mathbf{d}_i))$. After careful prompt engineering (see Table 3), the following prompt emerges as our primary choice: "`Provide an extremely specific question and short answer (one sentence) based on a fact present in the following document:` $\{\mathbf{d}_i\}$".

Quality Control In light of the above challenges, we first perform a careful post-processing to control the quality of the generated samples. (i) We remove instances

[3] Although ACE- ATTORNEY is model-agnostic, our infrastructure is restricted to 24 GB of GPU RAM. Thus, we can afford to run models up to 13B in size, also simulating a real-world hardware-constrained scenario.

Table 2 Two examples taken from SYN- LEQA featuring the root document and the corresponding generated question-answer pair

Document #1	
In August 2006, the EEOC brought this suit in the U.S. District Court for the District of New Jersey on behalf of two employees over aged 60 who received reduced severance pay after their plant closed. It alleged that reducing the severance pay of employees over 60 who qualified for pensions violated the Age Discrimination in Employment Act. On May 27, 2008, the court ruled in favor of the defendants and awarded no relief.	
QUESTION #1	Did the court rule in favor of the EEOC in the case of reduced severance pay for employees over 60 who qualified for pensions?
ANSWER #1	No, the court ruled in favor of the defendants and awarded no relief.
Document #2	
We may suspend or terminate your rights to use the Parsec Properties (including your Account) at any time for any reason at our sole discretion, including for any use of the Parsec Properties in violation of these Terms. Upon termination of your rights under these Terms, your Account and right to access and use the Parsec Properties will terminate immediately. In addition, we use "pixel tags" (also referred to as clear GIFs, Web beacons, or Web bugs). Pixel tags are tiny graphic images with a unique identifier, similar in function to cookies, which are used to track online movements of Web users.	
QUESTION #2	What is one specific reason that Parsec Properties may suspend or terminate a user's account and rights to use the Parsec Properties?
ANSWER #2	One specific reason that Parsec Properties may suspend or terminate a user's account and rights to use the Parsec Properties is for any use of the Parsec Properties in violation of these Terms.

Table 3 Prompts engineered for synthetic data generation (above) and zero-shot LLMs (below)

Setting	Prompt
Synthetic Data Generation	
`Task Description`†	Provide an extremely specific question and short answer (one sentence) based on a fact present on the following document: $\{d_i\}$. It should have the following format: Question: generated_question Answer: generated_answer.
`LLaMA Task Description`	<s>[INST]<<SYS>> You are an annotator that needs to provide an extremely specific question and short answer (one sentence) based on a fact present on a document. It should have the following format: Question: generated_question Answer: generated_answer. <</SYS>> The document is$\{d_i\}$ [/INST]
`In-Context Learning`	Relevant document: $\{relevant\_document_1\}$ Corresponding question and answer: Question: $\{question_1\}$ Answer: $\{answer_1\}$ Relevant document: $\{relevant\_document_2\}$ Corresponding question and answer: Question: $\{question_2\}$ Answer: $\{answer_2\}$ Relevant document: $\{d_i\}$ Corresponding question and answer:
`In-Context Learning w/ Task Description`	Provide an extremely specific question and short answer (one sentence) based on a fact present on the following document: $\{d_i\}$. It should have the following format: Question: generated_question Answer: generated_answer. Examples of questions and answers: Question: $\{question_1\}$ Question: $\{question_2\}$ Corresponding question and answer:
Zero-shot LLMs	
`Prompting`	Provide a single sentence answer to the following question: $\{q_i\}$. Using the following document: $\{d_i\}$
`Prompting w/ Selective Generative Paradigm`	Provide a single sentence answer to the following question: $\{q_i\}$. Using the following document: $\{d_i\}$ If the question is unanswerable given the provided document, write "Please provide the correct document."

† denotes the final prompt chosen for the creation of SYN- LEQA

with empty questions or answers. (ii) We address C2 by filtering samples with too generic questions through a retrieval-based approach. Technically, given $\mathbf{q}_i$, we exclude samples whose $\mathbf{d}_i$ is not retrieved in the first top-10 positions using BM25 OKAPI (Robertson et al. 1995) (more details are given in Section 4.3). Finally, to verify C3, we conduct a human analysis to gauge data quality. Specifically, we present 50 random samples to three English-proficient annotators with legal background asking to assign a score on a Likert scale from 1 (worst) to 5 (best), according to *fluency*, *factualness* (for both $\mathbf{q}_i$ and $\mathbf{a}_i$), and *informativeness* (for just $\mathbf{a}_i$). Figure 2 shows very high results, supported by inter-annotator agreement of 0.62 with Cohen's κ coefficient, revealing the great generative capacity of our LLM teacher. The evaluation criteria are detailed in Table 4.

3.2 Selective generative paradigm

The quality of the generated answer is highly dependent on the retriever $\mathcal{R}$, which prevents the generator $\mathcal{G}$ from producing a coherent and factual response when retrieving the wrong document. Consequently, previous work has put great effort into improving communication between $\mathcal{R}$ and $\mathcal{G}$, devising solutions that combine these two components in an end-to-end fashion (Lewis et al. 2021; Moro et al. 2022, 2023; Lai et al. 2023; Moro et al. 2023). However, $\mathcal{G}$ still places complete trust in $\mathcal{R}$, without having the possibility to replace the received document.

Inspired by communication networks that require adaptability to manage dynamic scenarios (Moro and Monti 2012), we advocate for a $\mathcal{G}$ capable of doubting the conduct of $\mathcal{R}$. We term this ability as *Selective Generative Paradigm* (SGP) (see Fig. 1), which allows $\mathcal{G}$ decide whether to generate the answer based on the current document retrieved or ask $\mathcal{R}$ to return a different one. Technically, if $\mathcal{G}$ is not satisfied with the selection, it must produce the following sentence **s**: `"Please provide the correct document."` Therefore, the second most ranked document is passed to $\mathcal{G}$, which re-assesses the relevance until it generates a proper answer. The formal procedure is given in Algorithm 1.

Regarding zero-shot LLMs, this approach is implemented by modifying the original prompt as described in Table 3. Differently, SLMs undergo an altered fine-tuning process (see the dataset refinement in Fig. 1). Specifically, we tamper the training set instances whose corresponding document is not correctly retrieved at the first

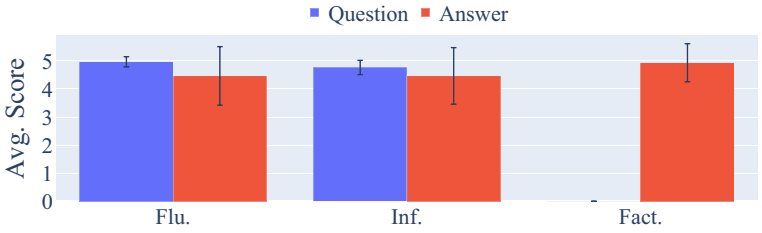


Fig. 2 Human analysis on SYN- LEQA's data quality according to the average *fluency*, *factualness*, and *informativeness* scores

Table 4 Explanations on human evaluation aspect scales

Informativeness	
1	Answer is not relevant to the question
2	Answer is partially relevant and misses the main point of the question
3	Answer is relevant, but misses the main point of the question
4	Answer successfully captures the main point of the question but some relevant content is missing
5	Answer successfully captures the main point of the question
Factualness	
1	Answer/question consists almost entirely of fabricated content that does not occur in the source document
2	Answer is mainly composed of hallucinations
3	Answer/question contains few hallucinations, but concern important aspects of the original document
4	Answer contains few hallucinations, but are restricted to negligible facts
5	Answer/question is faithful with respect to the original document
Fluency	
1	Answer/question is full of garbage fragments and is hard to understand
2	Answer/question contains fragments, missing components but has some fluent segments
3	Answer/question contains some grammar errors but is in general fluent
4	Answer/question has relatively minor grammatical errors
5	Fluent answer/question

Algorithm 1 Selective Generative Paradigm.

Input: $\mathbf{q}$ ▷ question
Output: $\mathbf{a}$ ▷ answer
1: $\{\mathbf{d}'_1, \ldots, \mathbf{d}'_n\} = \mathcal{R}(\mathbf{D} \mid \mathbf{q}),\ i = 1$
2: **while** $\mathbf{a} = \mathbf{s}$ **do**
3: $\quad \mathbf{a} = \mathcal{G}(\mathbf{q}, \mathbf{d}'_i)$
4: $\quad i = i + 1$
5: **end while**

position by BM25 OKAPI (Robertson et al. 1995) (see Section 4.3) by replacing their target answer with $\mathbf{s}$. Then, we fine-tune SLMs on this modified dataset, called SYNLEQA$_{sgp}$, whose altered training samples are 654 out of 3854.

4 Experimental setup

Based on Moro et al. (2018); Domeniconi et al. (2017), in addition to the in-domain experiment on the proposed SYN- LEQA dataset, we conduct a cross-domain evaluation on the PRIVACYQA dataset (Ravichander et al. 2019), aiming to assess the

performance of the distilled SLMs trained on SYN- LEQA. Since PRIVACYQA is centered on a classification task, we make adjustments to align it with our specific generative QA objective. In fact, in this dataset, for each query $\mathbf{q}_i$, expert annotators classified each sentence in the corresponding document $\mathbf{d}_i$ as relevant or not. Therefore, considering that there can be multiple sentences marked as relevant for $\mathbf{q}_i$, we select the longest as the candidate target answer $\mathbf{a}_i$. We call this testbed PRIVACYQA$_{gen}$ and outline its statistics in Table 1. Since the questions did not include enough context for the retriever to behave as expected, we do not consider $\mathcal{R}$ in the pipeline.

In summary, in the *in-domain scenario*, we use SGP and therefore train SLMs on SYN- LEQA$_{sgp}$. Differently, in the *cross-domain scenario*, we directly feed the gold document to both SLMs and LLMs without leveraging SGP. Hence, we train SLMs on the original SYN- LEQA dataset and test on PRIVACYQA$_{gen}$.

Environment All runs are tracked with Weights & Biases[4] and executed on a workstation with a single Nvidia GeForce RTX3090 GPU of 24 GB of dedicated memory, 64 GB of VRAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz. The reference operating system is Ubuntu 20.04.3 LTS. To enhance portability, our environment is built on top of a docker container with an NVIDIA image.[5] Our code is based on Python 3.8.10, PyTorch 1.12.1+cu113 (Paszke et al. 2019), and HuggingFace Transformers (Wolf et al. 2020).

4.1 Metrics

To thoroughly assess model performance, we cover multiple evaluation dimensions. Table 5 lists the hyperparameters of the metrics.

Syntactic We use recall-oriented **ROUGE-{1,2,L}** (Lin 2004) and precision-focused **BLEU-{1,2,3,4}** (Papineni et al. 2002). Inspired by Moro et al. (2023), we also measure an aggregated ROUGE judgment to penalize results with discrepant unigram, bigram, and longest common subsequence overlaps: $\mathcal{RG} = {}^{\text{avg}(r_1,r_2,r_L)}/_{1+\sigma_r^2}$, where σ_r^2 is the F1 variance. Accordingly, we apply the same principle to BLEU, obtaining $\mathcal{B} = {}^{\text{avg}(b_1,b_2,b_3,b_4)}/_{1+\sigma_b^2}$.

Semantic Moving beyond lexical superficiality, we employ **BERTScore** (BeS) (Zhang et al. 2020) and **BARTScore** (BaS) (Yuan et al. 2021) for semantic coverage, two model-based metrics that exhibit strong correlations with human judgment. Note that BARTScore computes the generation probability $p(\mathbf{y}|\mathbf{x}, \theta)$ of a sequence $\mathbf{y}$ conditioned on another sequence $\mathbf{x}$, where θ are the weights of a BART model (Lewis et al. 2020). Due to this generative approach, the evaluation dimensions vary depending on how $\mathbf{y}$ and $\mathbf{x}$ are defined. In particular, we consider the Recall, Precision and F1 settings. Recall ($\mathbf{h} \rightarrow \mathbf{r}$, $p(\mathbf{r}|\mathbf{h}, \theta)$) quantifies how easily a gold reference ($\mathbf{r}$) could be generated by the hypothesis ($\mathbf{h}$). Precision ($\mathbf{r} \rightarrow \mathbf{h}$, $p(\mathbf{h}|\mathbf{r}, \theta)$) assesses how likely the

[4] https://wandb.ai

[5] nvidia/cuda:11.3.1-devel-ubuntu20.04

Table 5 Hyperparameters initialization for utilized NLG metrics

Metric	Bound	Hyperparameters
BLEU	[0, 1]	`bleu_types=["bleu1", "bleu2", "bleu3", "bleu4"], weights="equidistributed"`
ROUGE	[0, 1]	`rouge_types=["rouge1","rouge2","rougeL"], use_aggregator=True, use_stemmer=True, metric_to_select="fmeasure"`
BERTScore	$[-1, 1]$	`model_type="bert-base-uncased", idf=True, batch_size=64, nthreads=4, rescale_with_baseline=True, use_fast_tokenizer=False, return_average_scores=False`
BARTScore	$]-\infty, 0[$	`model_checkpoint="facebook/bart-large-cnn", batch_size=4, segment_scores=False`

answer hypothesis could be constructed based on the gold reference. F1 score ($\mathbf{h} \leftrightarrow \mathbf{r}$) takes the harmonic mean of recall and precision.

Efficiency We monitor the **inference runtime** and compute **CO_2 emissions** with CodeCarbon.[6] Finally, we condense our judgment to weigh both cost and effectiveness in a single score using **Carburacy** (Moro et al. 2023).

4.2 Models

Figure 3 shows the models used in this study, specifying their checkpoints and hyperparameters for fine-tuning and inference.

SLMs In line with previous works (Ragazzi et al. 2024; Italiani et al. 2024), we consider **BART** (Lewis et al. 2020), one of the most popular models for generative tasks characterized by a denoising pretraining objective, and **FLAN-T5** (Chung et al. 2022), a model fine-tuned with instructions on a mixture of text-to-text tasks. We benchmark both models with different sizes, ranging from 60M to 780M parameters. For the fine-tuning process, we concatenate the question and the relevant document and train them to produce the corresponding answer. We train each model for 5 epochs and select the checkpoints that obtain the highest ROUGE-1 score on the validation set.

LLMs We examine **LLAMA-2-{7B,13B}** (Touvron et al. 2023a), which is an upgraded version of LLAMA- 1 (Touvron et al. 2023b) pretrained on additional 40% of data. **MISTRAL-7B** (Jiang et al. 2023) comprises grouped-query attention (Ainslie et al. 2023) and sliding window attention (Beltagy et al. 2020) to achieve a harmonious equilibrium between the pursuit of high performance and efficiency.

[6] https://github.com/mlco2/codecarbon

Model	Checkpoint
SLMs	
BART-base	facebook/bart-base
BART-large	facebook/bart-large
FLAN-T5-small	google/flan-t5-small
FLAN-T5-base	google/flan-t5-base
FLAN-T5-large	google/flan-t5-large
LLMs	
LLAMA-2-7B	meta-llama/Llama-2-7b-chat-hf
LLAMA-2-13B	meta-llama/Llama-2-13b-chat-hf
MISTRAL-7B	mistralai/Mistral-7B-Instruct-v0.1
ZEPHYR-7B	HuggingFaceH4/zephyr-7b-alpha
ORCA-2-7B	microsoft/Orca-2-7b
ORCA-2-13B	microsoft/Orca-2-13b
NEURAL-CHAT-7B	Intel/neural-chat-7b-v3-1
STARLING-7B	berkeley-nest/Starling-LM-7B-alpha
NOTUS-7B	argilla/notus-7b-v1

Hyperparameter	Value
Dropout rate†	0.1
Learning rate†	$5e^{-5}$, linear scheduler
AdamW Opt. [54]†	0.9 β_1, 0.999 β_2, $1e^{-2}$ weight decay
Batch size†	2
Epochs†	5 (validation every epoch)
Decoding strategy	greedy search, max_length=300
Seed	42
Temperature	1.0
Quantization‡	8-bit

Fig. 3 *Left*: HuggingFace checkpoints of the models used in this study. *Right*: Hyperparameters used for SLMs fine-tuning and inference, and zero-shot LLMs. † and ‡ are specific for SLMs and LLMs, respectively

ZEPHYR-7B (Tunstall et al. 2023) leverages MISTRAL- 7B as a starting checkpoint, distilling conversational capabilities with direct preference optimization from a dataset of AI feedback. **ORCA-2-{7B,13B}** (Mitra et al. 2023) is a refined iteration of LLAMA- 2 that outperforms models of comparable scale on tasks requiring advanced reasoning abilities. **NEURAL-CHAT-7B** is a model based on MISTRAL- 7B fine-tuned on SLIMORCA. **STARLING-7B** (Zhu et al. 2023) is trained with reinforcement learning from AI harnessing NECTAR, a GPT-4 labeled ranking dataset composed of chat prompts. Lastly, through fine-tuning on a refined variant of ULTRAFEEDBACK (Cui et al. 2023), **NOTUS-7B** (Bartolome et al. 2023) elevates the performance of ZEPHYR-7B. To make inference for zero-shot LLMs possible with our hardware, we exploit PEFT.[7] To be precise, we carry out 8-bit model quantization, enclosing the provided prompt within the recommended template for every model.

4.3 Retriever

The retriever is a pivotal component in real-world QA pipelines, having a direct influence on the overall accuracy, efficiency, and reliability of the entire solution (Frisoni et al. 2022). In light of this, we perform a comparison of different methods with the aim of maximizing performance while minimizing costs. First, we test **BM25 OKAPI** (Robertson et al. 1995), a probabilistic model that estimates the relevancy of documents w.r.t. a query as follows:

$$\text{BM25} = \sum_{t \in q} \log\left(\frac{n - f_t + 0.5}{f_t + 0.5}\right) \cdot \text{TF}_{\text{BM25}} \tag{1}$$

$$\text{TF} = \frac{f_{t,\mathbf{d}} \cdot (k_1 + 1)}{f_{t,\mathbf{d}} + k_1 \cdot \left((1 - b) + \left(b \cdot \ell_{\mathbf{d}} / \ell_{\text{avg}}\right)\right)} \tag{2}$$

[7] https://github.com/huggingface/peft

where n is the number of documents in the corpus, f_t is the number of documents containing the query term t, $f_{t,\mathbf{d}}$ is the number of occurrences of t in the document $\mathbf{d}$, $\ell_{\mathbf{d}}$ is the number of terms in $\mathbf{d}$, and ℓ_{avg} is the average of $\ell_{\mathbf{d}}$ over the corpus. We use the standard parameters: $k_1 = 1.2$ and $b = 0.75$.

Furthermore, we evaluate the *bi-encoder* (BiE) architecture (Reimers and Gurevych 2019), which uses a siamese BERT network to create semantically meaningful sentence embeddings. This involves independently processing $\mathbf{q}$ and $\mathbf{d}_i$ through an encoder-only transformer, resulting in two unique sentence embeddings that can be compared using cosine similarity (Frisoni et al. 2020; Domeniconi et al. 2016). To enhance computational efficiency, we conduct indexing using FAISS (Johnson et al. 2019), which tackles the problem of efficient comparison between high-dimensional vectors. Lastly, we assess the *cross-encoder* (CE) implementation (Nogueira and Cho 2019), using an encoder-only transformer to process the concatenated $\mathbf{q}$ and $\mathbf{d}_i$, producing a score indicating the document pertinence. Looking at the MS Marco Passage Reranking (Nguyen et al. 2016) and TREC (Craswell et al. 2020) benchmarks, we operate **MPNET** (Song et al. 2020) and **DISTILROBERTA** (Sanh et al. 2019) as bi-encoders, and **TINYBERT** (Jiao et al. 2020) and **MINILM** (Wang et al. 2020) as cross-encoders.

Results Table 6 compares the performance of different retrieval methods on the test set of SYN- LEQA. We report the Rank@k metric, with $k \in \{1, 5, 10\}$. Technically, given an input question, we assess how frequently the corresponding document is correctly found within the top-k items retrieved. We also provide the runtime in seconds and the CO_2 emissions in milligrams. According to Reimers and Gurevych (2019), cross-encoders exhibit the best results at the price of considerably higher computational costs because they require computing a score for every possible combination of query and document. This assertion is also validated by our case study. For example, although CE MINILM demonstrates superior performance w.r.t. BM25 OKAPI (90.06 vs. 83.02 for Rank@1), it requires more than 100x runtime (1428.88 vs. 10.89 seconds) and emissions (162.58 vs. 1.09 milligrams). Additionally, the performance gap decreases sharply as k increases (97.52 vs. 95.65 for Rank@5). For these reasons, to strike the best equilibrium between performance and computational overhead, we select BM25 OKAPI as the retriever for our in-domain experiments.

Table 6 Retrieval results

Model	R@1 ↑	R@5 ↑	R@10 ↑	Run. ↓	Emis. ↓
BM25 OKAPI	83.02	95.65	97.10	**10.89**	**1.09**
BiE DISTILROBERTA	70.81	87.78	90.89	<u>12.20</u>	<u>1.44</u>
BiE MPNET	73.50	87.16	91.72	18.02	2.15
CE TINYBERT	<u>89.65</u>	<u>96.80</u>	<u>98.14</u>	1473.71	32.15
CE MINILM	**90.06**	**97.52**	**98.76**	1428.88	162.58

Bold and underline denote the best and second-best scores. ↑=higher is better; ↓=lower is better

5 Results and discussion

Prompt Engineering As emphasized in previous studies (Khashabi et al. 2022; Wei et al. 2022; Ye et al. 2022), the configuration of the prompt template can profoundly influence the process of synthetic data generation. Therefore, we select 100 documents equally drawn from lawsuits and policies. From these documents, we generate four different versions of SYN- LEQA, with each variant differing based on the specific approach used to generate the synthetic questions and answers. These variations are produced by applying four distinct prompt engineering strategies: (see Table 3): (i) *Task Description*, we provide a textual description of the task and a test example; (ii) *In-Context Learning*, the model is given a few demonstrations of the task at inference time, but no weight updates are allowed; (iii) LLAMA *Task Description*, we include the same special tokens used during LLAMA- 2 training with a wide variety of system prompts intended for different tasks; (iv) *In-Context Learning with Task Description*. We then fine-tune FLAN- T5-large on these four different reduced SYN- LEQA's versions and test it on PRIVACYQA$_{gen}$. With this experiment, we aim to test which prompt engineering technique generates the most effective synthetic dataset, ultimately leading to optimal cross-domain performance on PRIVACYQA for a model fine-tuned on that data. Table 7 indicates that *Task Description* performs the best, while LLAMA *Task Description* exhibits slightly higher efficiency, albeit with reduced effectiveness compared to the former. As expected, strategies that use in-context learning take longer to generate question–answer pairs because the model is provided with lengthier instructions as input. Moreover, this appears to harm the data generation process, resulting in extensively inferior results compared to methods that abstain from in-context learning. Consequently, all experiments are conducted with the prompt template of *Task Description*.

Effectiveness Table 8 shows the in-domain results on SYN- LEQA using SGP. Across both syntactic and semantic metrics, distilled SLMs consistently outperform LLMs. The greatest improvement is registered compared to LLAMA- 2- 13B, ORCA- 2- 13B, and MISTRAL- 7B. The performance gap is reduced compared to the newly introduced STARLING- 7B, ZEPHYR- 7B, and NEURAL- CHAT- 7B. Of particular interest are the enhanced retrieval results obtained thanks to SGP. Specifically, unlike LLMs, SLMs excel at identifying documents that have not been retrieved properly. The ability of the models to classify the retrieved documents as relevant or not is summarized in Table 9.

Table 7 Results of FLAN- T5-large on PRIVACYQA$_{gen}$'s test set fine-tuned on 100 instances from four versions of SYN- LEQA, each generated using different prompt templates

Prompt	$\mathcal{RG}$	BeS	Run.
`In-context learning`	21.47	24.74	10672
`Task description`	**30.88**	**33.92**	<u>1591</u>
`LLaMA task description`	<u>25.48</u>	<u>28.44</u>	**1462**
`In-context learn. w/ task descr.`	22.86	25.87	5061

Bold and underline denote the best and second-best results

Table 8 Quantitative results on SYN- LEQA's test set

SYN- LEQA													
			Syntactic					Semantic		Efficiency			
Model	Size	Release	$\mathcal{B}$	R-1	R-2	R-L	$\mathcal{RG}$	BaS-F1	BeS	Carb.	Run.	Emis.	R@1
`Fine-tuned SLMs`													
FLAN- T5- small	60M	`2022/10`	44.09	56.31	44.00	56.02	51.94	-2.15	55.86	75.14	195.49	6.66	75.78
BART- base	149M	`2020/07`	46.80	62.66	50.31	61.96	58.12	-1.92	62.21	78.88	161.17	7.42	88.61
FLAN- T5- base	250M	`2022/10`	49.70	63.48	50.68	62.96	58.83	-1.84	64.33	79.28	240.16	8.43	84.47
BART- large	406M	`2020/07`	52.51	65.85	53.77	65.34	61.46	-1.79	66.06	80.78	249.15	9.86	89.03
FLAN- T5- large	780M	`2022/10`	**53.37***	**66.96***	**54.70***	**65.88***	**62.32***	**-1.70***	**67.59***	**81.07***	481.16*	21.61*	**90.48***
`Zero-shot LLMs`													
LLAMA- 2	7B	`2023/07`	36.38	50.39	37.95	49.15	45.69	-2.78	41.72	69.74	2390.76	98.79	82.61
MISTRAL	7B	`2023/09`	12.76	41.94	29.56	41.17	37.44	-2.83	36.79	59.01	6650.08	512.61	54.04
ZEPHYR	7B	`2023/10`	16.45	57.56	43.00	54.49	51.48	-1.92	50.84	63.36	7070.25	883.53	82.19
ORCA- 2	7B	`2023/11`	22.41	50.05	36.32	48.29	44.72	-2.66	36.09	64.98	5814.93	410.86	81.37
NEURAL- CHAT	7B	`2023/11`	25.18	54.11	34.76	49.77	45.90	-2.64	45.67	68.15	4346.67	223.94	82.19
STARLING	7B	`2023/11`	33.13	58.69	45.08	56.57	53.26	-2.43	54.32	73.45	4687.27	172.07	82.19
NOTUS	7B	`2023/11`	13.16	50.03	36.33	48.21	44.69	-2.12	46.08	42.20	30687.14	3237.09	77.23
LLAMA- 2	13B	`2023/07`	35.59	45.77	33.52	44.05	40.99	-2.99	35.41	64.86	3871.89	225.12	80.33
ORCA- 2	13B	`2023/11`	12.21	43.04	30.12	41.84	38.20	-2.49	33.85	59.36	14176.05	527.61	77.85

Bold and underline denote the best and second-best results. * indicates statistically significant higher scores compared to the best-performing LLM (STARLING) based on a one-tailed t-test (p-value= 0.05)

Table 9 Classification results of SGP

Model	Size	Precision	Recall	F1	Accuracy
`Fine-tuned SLMs`					
FLAN- T5- small	60M	56.63	57.32	56.97	85.30
BART-base	149M	<u>86.30</u>	<u>76.83</u>	81.29	<u>94.00</u>
FLAN- T5-base	250M	79.17	69.51	74.03	91.72
BART-large	406M	84.72	74.39	**96.06**	93.37
FLAN- T5-large	780M	**86.67**	**79.27**	<u>82.80</u>	**94.41**
`Zero-shot LLMs`					
LLAMA- 2	7B	0.00	0.00	0.00	82.40
MISTRAL	7B	20.79	51.22	29.58	58.59
ZEPHYR	7B	68.42	15.85	25.74	84.44
ORCA- 2	7B	61.90	15.85	25.24	84.06
NEURAL- CHAT	7B	85.19	28.05	42.20	86.96
STARLING	7B	78.95	18.29	29.70	85.30
NOTUS	7B	48.15	31.71	38.24	82.61
LLAMA- 2	13B	27.03	12.20	16.81	79.50
ORCA- 2	13B	53.70	35.37	42.65	83.85

Bold and underline denote the best and second-best results. Following standard practice (Frisoni and Moro 2021), we evaluate model performance using precision, recall, F1-score, and accuracy metrics

Technically, we label the generated answer as 1 when the model outputs **s** (indicating an incorrectly retrieved document) and 0 otherwise. These findings show the benefits of SGP, which on average solidly increases the original R@1 score of 83.02, as shown in Table 6. The enhanced retrieval performance of SLMs also explains their superior question-answering effectiveness w.r.t. LLAMA- 2- 13B. Since this LLM was used to produce SYN- LEQA, superior results would be expected when evaluated on the same questions it generated. However, its performance is hindered by a markedly decreased R@1 score, which ultimately results in less informed answers. Furthermore, Table 10 shows the cross-domain results on PRIVACYQA$_{gen}$, where we test SLMs trained on SYN- LEQA. Again, SLMs overall outperform the majority of LLMs, albeit with a less pronounced improvement. First, this occurs because we transfer the question-answering capabilities of the models, acquired on SYN- LEQA, to a completely different dataset. Second, this setting does not include the retrieval module. In scenarios where the retrieval component is not needed, SLMs cannot benefit from SGP, hence the improvement w.r.t. LLMs shrinks. Moreover, we highlight that even when the best-performing SLM underperforms compared to the top LLM, the difference is not statistically significant except for R-1. Figure 4 visually summarizes the results.

Efficiency In addition to the better effectiveness of distilled SLMs, from Table 8 and Table 10, we can also appreciate the radical reduction in runtime and carbon emissions at inference time. For example, the best-performing LLM on PRIVACYQA$_{gen}$, such as STARLING- 7B, is characterized by $\approx$1200% increment in CO_2 emissions compared

Table 10 Quantitative results on the test set of PRIVACYQA$_{gen}$

PRIVACYQA$_{gen}$			Syntactic					Semantic		Efficiency		
Model	Size	Release	$\mathcal{B}$	R-1	R-2	R-L	$\mathcal{RG}$	BaS-F1	BeS	Carb.	Run.	Emis.
`Fine-tuned SLMs`												
FLAN- T5- small	60M	`2022/10`	17.21	30.65	19.68	32.01	27.36	-3.34	30.54	56.86	102.12	7.54
BART- base	139M	`2020/07`	11.40	28.31	16.29	29.85	24.73	-3.45	28.71	54.44	80.47	5.70
FLAN- T5- base	250M	`2022/10`	13.12	29.68	19.13	31.31	26.63	-3.32	29.92	56.19	127.07	7.83
BART- large	406M	`2020/07`	18.03	33.04	20.70	34.10	29.17	-3.26	32.92	58.40	160.82	11.74
FLAN- T5- large	780M	`2022/10`	18.00	35.16	**23.20***	35.90	**31.31**	-3.15*	34.55	**60.19***	284.08*	14.18*
`Zero-shot LLMs`												
LLAMA- 2	7B	`2023/07`	9.18	27.94	12.45	27.56	22.53	-3.62	27.95	50.80	1693.14	131.46
MISTRAL	7B	`2023/09`	12.90	29.08	14.31	29.21	24.08	-3.52	28.39	51.06	2497.99	238.00
ZEPHYR	7B	`2023/10`	11.57	37.65	19.43	35.34	30.61	**-2.85**	31.29	52.52	3683.45	597.87
ORCA- 2	7B	`2023/11`	11.07	26.39	8.78	24.34	19.71	-4.03	14.85	47.21	2632.77	199.24
NEURAL- CHAT	7B	`2023/11`	5.43	24.32	6.02	22.66	17.55	-4.09	21.24	44.32	2282.09	257.43
STARLING	7B	`2023/11`	**19.83**	**38.22***	19.64	**36.30**	31.17	-3.63	**34.66**	57.81	3220.92	183.61
NOTUS	7B	`2023/11`	11.12	36.99	16.70	33.06	28.70	-3.01	32.38	41.81	11047.64	1684.49
LLAMA- 2	13B	`2023/07`	12.07	27.65	12.75	27.59	22.55	-3.56	27.54	50.55	2823.87	155.35
ORCA- 2	13B	`2023/11`	7.28	25.61	9.89	24.26	19.82	-3.40	19.93	41.83	13044.20	791.71

Bold and underline denote the best and second-best results. * indicates statistically significant higher performance compared to the best-performing LLM (STARLING) based on a one-tailed t-test (p-value= 0.05)

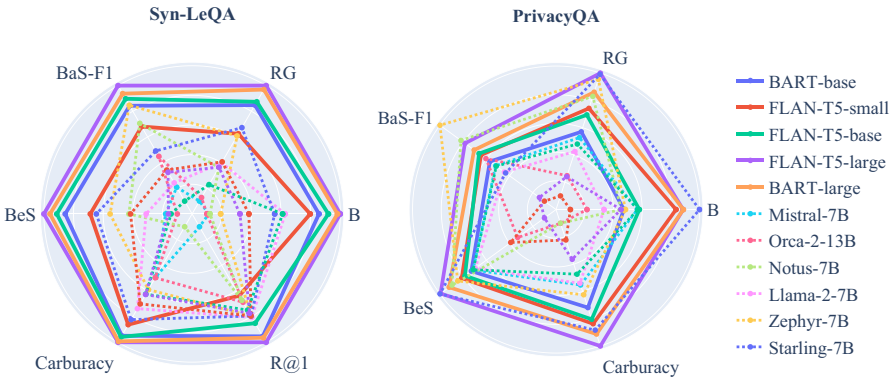


Fig. 4 Graphical overview of model effectiveness, where dotted lines represent LLMs and solid lines SLMs

to the best SLM, such as FLAN- T5-large. Figure 5 illustrates the relationship between efficiency and effectiveness, revealing the high performance of SLMs with low costs.

Few-shot Setting Limitations due to low-resource regimes, such as commodity hardware infrastructure, can affect model performance (Parida and Motlícek 2019; Moro and Ragazzi 2022, 2023; Huh and Ko 2023; Moro et al. 2023). Regardless, it is plausible that a real-world organization lacks adequate resources to generate thousands of

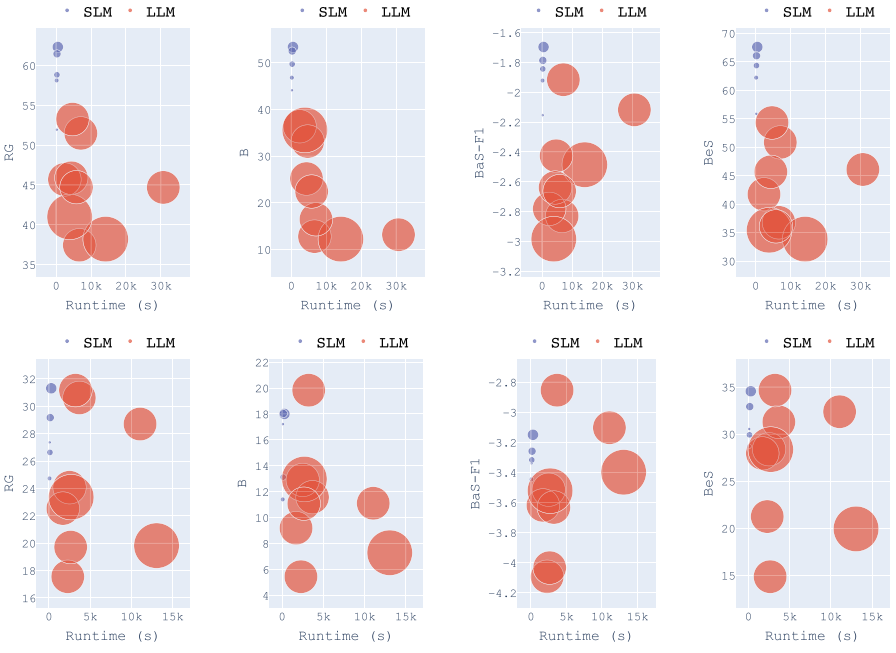


Fig. 5 Performance–efficiency relation. Top: SYN- LEQA; bottom: PRIVACYQA$_{gen}$. The size of bubbles refers to the number of model parameters

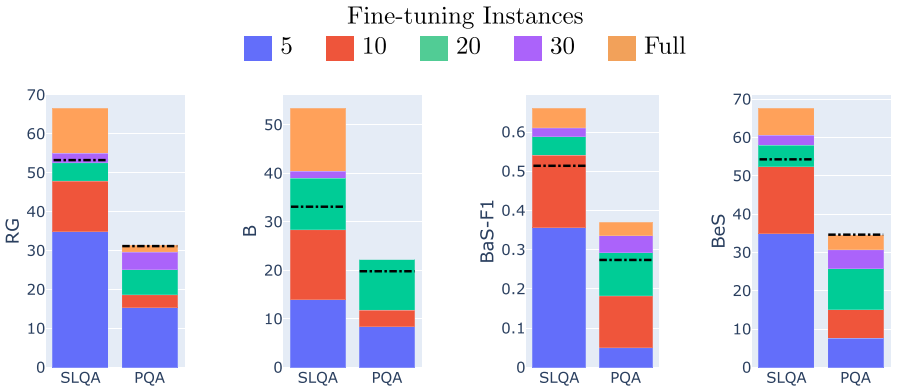
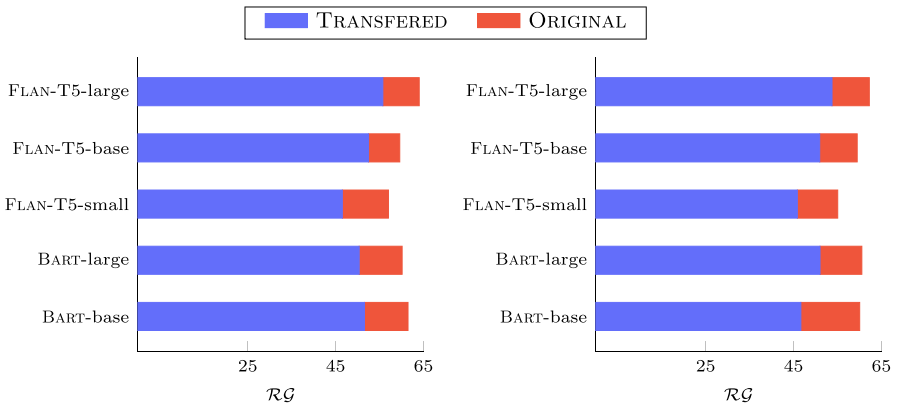


Fig. 6 Few-shot setting results for SYN- LEQA (SLQA) and PRIVACYQA$_{gen}$ (PQA). The dashed black bands mark the top-performing LLM (STARLING- 7B)

synthetic examples. Additionally, there might be a need for a thorough examination of the created question–answer pairs before integrating them as references for the production model. Accordingly, it is of particular interest to study the performance of the student model w.r.t. the number of synthetic training instances available. Figure 6 summarizes the results of FLAN- T5-large fine-tuned on an increasing number of SYN- LEQA samples, drawn equally from lawsuits and policies. Notably, in SYN- LEQA, the SLM surpasses the best-performing LLM, such as STARLING- 7B, with only 20 labeled samples. In PRIVACYQA$_{gen}$, we notice the same behavior in 2 out of 4 metrics, such as BLEU and BARTScore.

Transfer Learning We evaluate the transferability of SLMs by assessing their performance when trained on one subset of SYN- LEQA (i.e., lawsuits or policies) and



(a) Policies → Lawsuits (b) Lawsuits → Policies

Fig. 7 Transferability results on SYN- LEQA. TRANSFERED denotes a model trained on one subset and tested on the other. ORIGINAL denotes a model trained and tested on the same subset

Table 11 Flan-T5-large performance on the test set of PRIVACYQA$_{gen}$

Training Set	$\mathcal{B}$	R-1	R-2	R-L	$\mathcal{RG}$	BaS-F1	BeS
PRIVACYQA$_{gen}$	**80.68**	85.36	82.33	85.52	84.39	-0.972	85.53
SYN- LEQA + PRIVACYQA$_{gen}$	80.49	**90.10**	**88.35**	**90.58**	**89.67**	**-0.717**	**90.43**

Bold entries underline the best results

tested on the other. We report the results with the same subset for training and testing as upper bound. Figure 7 demonstrates the adaptability of SLMs, showcasing consistent ROUGE performance when transferred between different domains. Additionally, we perform an experiment to assess the extent to which a model can leverage knowledge acquired during its initial training on SYN- LEQA and transfer it to improve performance on PRIVACYQA$_{gen}$. Specifically, we first train the best-performing model, FLAN- T5-large, on our synthetic dataset, and then fine-tune it on PRIVACYQA$_{gen}$. For this purpose, we divide the original PRIVACYQA$_{gen}$ dataset, which originally contains only a test set, into three subsets: 344 instances for training, 43 for validation, and 44 for testing. Table 11 demonstrates that the model fine-tuned solely on PRIVACYQA$_{gen}$ performs worse than when it is initially fine-tuned on SYN- LEQA, underscoring the usefulness of our synthetic dataset.

6 Conclusion

In this paper, we introduce ACE- ATTORNEY, an approach designed to produce LQA-specific datasets and supervised SLMs through LLM KD. Precisely, given an input textual prompt, a frozen LLM generates artificial samples that are used as knowledge to train a reduced-parameter student model. We collect LLM-crafted samples and create SYN- LEQA, our synthetic dataset proposed to address the scarcity of public LQA corpora. Experiments carried out on both SYN- LEQA and a real expert-labeled dataset showcase the ability of student models to surpass the latest cutting-edge LLMs. Importantly, distilled models operate with amply fewer parameters and demand considerably shorter inference runtime and CO_2 emissions. Moreover, simulating a real-world retrieval-based scenario, we show that the Selective Generative Paradigm enables the retriever and the generator to better select the correct document to answer a given question. Finally, we reveal that our specialized SLMs achieve competitive performance even when we have access to a limited number of synthetic instances, allowing manual oversight at a reasonable cost across the entire generated dataset.

Ethics Statement and limitations

Our research holds noteworthy promise for advancing the development of new LQA datasets and models. However, given the widely recognized issue of hallucinations in LLMs, experts must examine the authenticity and reliability of the generated data. In fact, our aim is not to demonstrate that LLMs could replace legal professionals. Instead,

we introduce ACE- ATTORNEY as a research instrument, highlighting its complementary function rather than positioning it as a substitute for detailed and contextually specific efforts.

Despite the large amount of up-to-date LLMs used in our experiments, we focus on models capable of working within the constraints of a single 24 GB of GPU RAM. Consequently, this excluded larger and potentially more performant solutions from our study, which could affect the synthetic data generation process. However, human examination suggested high quality despite the use of a 13B LLM. Furthermore, the cross-domain scope of our study regarding the evaluation of our fine-tuned SLMs on real human-annotated datasets is constrained, as we considered only a single dataset in our analysis. However, this decision is due to PRIVACYQA$_{gen}$ being the only dataset that meets the necessary criteria to act as a benchmark for generative LQA. Finally, due to the limited availability of multilingual legal datasets (Niklaus et al. 2023) and solutions (Moro et al. 2024a), future directions should prioritize advances in low-resource languages with a focus on interpretable responses (Moro et al. 2024).

Acknowledgements This research is partially supported by (i) Artificial Intelligence for Public Administration Connected (AI-PACT), CUP B47H22004450008 and B47H22004460001, PNRR, mission 4, component 2, investment 2.3, (ii) the Complementary National Plan PNC-I.1, "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE—DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (iii) the PNRR, M4C2, FAIR—Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program. We thank the Maggioli Group (https://www.maggioli.com/who-we-are/company-profile) for partially supporting the Ph.D. scholarship granted to Paolo Italiani.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Ahmad WU, Chi J, Tian Y, Chang K (2020) Policyqa: A reading comprehension dataset for privacy policies. In: Cohn T, He Y, Liu Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event. Findings of ACL, vol. EMNLP 2020, pp 743–749. Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.66

Ainslie J, Lee-Thorp J, Jong M, Zemlyanskiy Y, Lebrón F, Sanghai S (2023) Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245

Almazrouei E, Alobeidli H, Alshamsi A, Cappelli A, Cojocaru R, Debbah M, Goffinet É, Hesslow D, Launay J, Malartic Q, Mazzotta D, Noune B, Pannier B, Penedo G (2023) The falcon series of open language models. CoRR arxiv:2311.16867. https://doi.org/10.48550/ARXIV.2311.16867

Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, Tepper N, Zwerdling N (2020) Do not have enough data? deep learning to the rescue! In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 7383–7390

Bartolome A, Martin G, Vila D (2023) Notus. GitHub

Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150

Bird S, Klein E, Loper E (2017) Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc

Bongard L, Held L, Habernal I (2022) The legal argument reasoning task in civil procedure. In: Aletras N, Chalkidis I, Barrett L, Goanta C, Preotiuc-Pietro D (eds) Proceedings of the Natural Legal Language Processing Workshop, NLLP@EMNLP 2022, Abu Dhabi, United Arab Emirates (Hybrid), pp 194–207. Association for Computational Linguistics. https://aclanthology.org/2022.nllp-1.17

Bonifacio L, Abonizio H, Fadaee M, Nogueira R (2022) Inpars: Data augmentation for information retrieval using large language models. arXiv preprint arXiv:2202.05144

Carranza AG, Farahani R, Ponomareva N, Kurakin A, Jagielski M, Nasr M (2023) Privacy-preserving recommender systems with synthetic query generation using differentially private large language models. arXiv preprint arXiv:2305.05973

Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li E, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Narang S, Mishra G, Yu A, Zhao VY, Huang Y, Dai AM, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J (2022) Scaling instruction-finetuned language models. CoRR arxiv:2210.11416. https://doi.org/10.48550/ARXIV.2210.11416

Craswell N, Mitra B, Yilmaz E, Campos D, Voorhees EM (2020) Overview of the TREC 2019 deep learning track. CoRR arxiv:2003.07820

Cui G, Yuan L, Ding N, Yao G, Zhu W, Ni Y, Xie G, Liu Z, Sun M (2023) Ultrafeedback: Boosting language models with high-quality feedback. CoRR arxiv:2310.01377. https://doi.org/10.48550/ARXIV.2310.01377

Domeniconi G, Moro G, Pagliarani A, Pasolini R et al (2017) On deep learning in cross-domain sentiment classification. In: Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-(Volume 1), vol 1, pp 50–60. SciTePress

Domeniconi G, Moro G, Pasolini R, Sartori C (2014) Cross-domain text classification through iterative refining of target categories representations. In: International Conference on Knowledge Discovery and Information Retrieval, vol 2, pp 31–42. SciTePress

Domeniconi G, Moro G, Pasolini R, Sartori C (2015) Iterative refining of category profiles for nearest centroid cross-domain text classification. In: Knowledge Discovery, Knowledge Engineering and Knowledge Management: 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers 6, pp 50–67. Springer

Domeniconi G, Semertzidis K, Lopez V, Daly EM, Kotoulas S, Moro G (2016) A novel method for unsupervised and supervised conversational message thread detection. In: International Conference on Data Management Technologies and Applications, vol 2, pp 43–54. SciTePress

Faggioli G, Dietz L, Clarke CL, Demartini G, Hagen M, Hauff C, Kando N, Kanoulas E, Potthast M, Stein B et al (2023) Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp 39–50

Frisoni G, Mizutani M, Moro G, Valgimigli L (2022) Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 5770–5793

Frisoni G, Moro G (2021) Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In: Data Management Technologies and Applications: 9th International Conference, DATA 2020, Virtual Event, July 7–9, 2020, Revised Selected Papers 9, pp 293–318. Springer

Frisoni G, Moro G, Carbonaro A et al (2020) Learning interpretable and statistically significant knowledge from unlabeled corpora of social text messages: A novel methodology of descriptive text mining. In: DATA, pp 121–132

Gao J, Pi R, Lin Y, Xu H, Ye J, Wu Z, Zhang W, Liang X, Li Z, Kong L (2023) Self-guided noise-free data generation for efficient zero-shot learning. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda. OpenReview.net. https://openreview.net/pdf?id=h5OpjGd_lo6

Ghosh S, Evuru CKR, Kumar S, S R, Sakshi S, Tyagi U, Manocha D (2023) DALE: generative data augmentation for low-resource legal NLP. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, pp

8511–8565. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.528

Guo Z, Wang P, Wang Y, Yu S (2023) Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. arXiv preprint arXiv:2305.07804

Hinton GE, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. CoRR arxiv:1503.02531

Holzenberger N, Blair-Stanek A, Durme BV (2020) A dataset for statutory reasoning in tax law entailment and question answering. In: Aletras N, Androutsopoulos I, Barrett L, Meyers A, Preotiuc-Pietro D (eds) Proceedings of the Natural Legal Language Processing Workshop 2020 Co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), Virtual Workshop. CEUR Workshop Proceedings, vol 2645, pp 31–38. CEUR-WS.org. https://ceur-ws.org/Vol-2645/paper5.pdf

Huh T, Ko Y (2023) Efficient framework for low-resource abstractive summarization by meta-transfer learning and pointer-generator networks. Expert Syst Appl 234:121029. https://doi.org/10.1016/J.ESWA.2023.121029

Italiani P, Frisoni G, Moro G, Carbonaro A, Sartori C (2024) Evidence, my dear watson: Abstractive dialogue summarization on learnable relevant utterances. Neurocomputing 572:127132

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, Bressand F, Lengyel G, Lample G, Saulnier L et al (2023) Mistral 7b. arXiv preprint arXiv:2310.06825

Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q (2020) Tinybert: Distilling BERT for natural language understanding. In: Cohn T, He Y, Liu Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Findings of ACL, vol. EMNLP 2020, pp 4163–4174. Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.372

Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with gpus. IEEE Trans Big Data 7(3):535–547

Khashabi D, Baral C, Choi Y, Hajishirzi H (2022) Reframing instructional prompts to gptk's language. In: Muresan S, Nakov P, Villavicencio A (eds) Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pp 589–612. Association for Computational Linguistics. https://doi.org/10.18653/V1/2022.FINDINGS-ACL.50

Kim M-Y, Goebel R, Kano Y, Satoh K (2016) Coliee-2016: evaluation of the competition on legal information extraction and entailment. In: International Workshop on Juris-informatics (JURISIN 2016)

Kumar V, Choudhary A, Cho E (2020) Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245

Lai TM, Castellucci G, Kuzi S, Ji H, Rokhlenko O (2023) External knowledge acquisition for end-to-end document-oriented dialog systems. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp 3633–3647

Lee K, Guu K, He L, Dozat T, Chung HW (2021) Neural data augmentation via example extrapolation. arXiv preprint arXiv:2102.01335

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp 7871–7880. Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.ACL-MAIN.703

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Lin C-Y (2004) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp 74–81. Association for Computational Linguistics, Barcelona, Spain. https://aclanthology.org/W04-1013

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. In: ICLR. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

Louis A, Dijck G, Spanakis G (2023) Interpretable long-form legal question answering with retrieval-augmented large language models. CoRR arxiv:2309.17050. https://doi.org/10.48550/ARXIV.2309.17050

MacAvaney S, Soldaini L (2023) One-shot labeling for automatic relevance estimation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2230–2235

Martinez-Gil J (2023) A survey on legal question–answering systems. Comp Sci Rev 48:100552. https://doi.org/10.1016/j.cosrev.2023.100552

Meng Y, Huang J, Zhang Y, Han J (2022) Generating training data with language models: Towards zero-shot language understanding. Adv Neural Inf Process Syst 35:462–477

Mitra A, Corro LD, Mahajan S, Codas A, Simões C, Agrawal S, Chen X, Razdaibiedina A, Jones E, Aggarwal K, Palangi H, Zheng G, Rosset C, Khanpour H, Awadallah AH (2023) Orca 2: Teaching small language models how to reason. CoRR **abs/2311.11045**. https://doi.org/10.48550/ARXIV.2311.11045

Moro G, Monti G (2012) W-grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. J Netw Comput Appl 35(4):1218–1234

Moro G, Ragazzi L (2023) Align-then-abstract representation learning for low-resource summarization. Neurocomputing 548:126356. https://doi.org/10.1016/J.NEUCOM.2023.126356

Moro G, Ragazzi L, Valgimigli L, Frisoni G, Sartori C, Marfia G (2023) Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. Sensors 23(7):3542. https://doi.org/10.3390/S23073542

Moro G, Piscaglia N, Ragazzi L, Italiani P (2024) Multi-language transfer learning for low-resource legal case summarization. Artif Intell Law 32(4):1111–1139. https://doi.org/10.1007/S10506-023-09373-8

Moro G, Pagliarani A, Pasolini R, Sartori C et al (2018) Cross-domain & in-domain sentiment analysis with memory-based deep neural networks. In: Proceedings of the 10th International Joint Conference on KnowledgeDiscovery, Knowledge Engineering and Knowledge Management, IC3K2018, Volume 1: KDIR, Seville, Spain, September 18-20, 2018, pp 125–136. SciTePress

Moro G, Ragazzi L (2022) Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp 11085–11093. AAAI Press. https://doi.org/10.1609/AAAI.V36I10.21357

Moro G, Ragazzi L, Valgimigli L (2023) Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In: Williams B, Chen Y, Neville J (eds) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pp. 14417–14425. AAAI Press. https://doi.org/10.1609/AAAI.V37I12.26686 . https://doi.org/10.1609/aaai.v37i12.26686

Moro G, Ragazzi L, Valgimigli L (2023) Graph-based abstractive summarization of extracted essential knowledge for low-resource scenarios. In: Gal K, Nowé A, Nalepa GJ, Fairstein R, Radulescu R (eds) ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023). Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 1747–1754. IOS Press. https://doi.org/10.3233/FAIA230460 . https://doi.org/10.3233/FAIA230460

Moro G, Ragazzi L, Valgimigli L, Freddi D (2022) Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, pp 180–189. Association for Computational Linguistics. https://doi.org/10.18653/V1/2022.ACL-LONG.15

Moro G, Ragazzi L, Valgimigli L, Molfetta L (2023) Retrieve-and-rank end-to-end summarization of biomedical studies. In: Pedreira O, Estivill-Castro V (eds) Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, Proceedings. Lecture Notes in Computer Science, vol 14289, pp 64–78. Springer. https://doi.org/10.1007/978-3-031-46994-7_6

Moro G, Ragazzi L, Valgimigli L, Vincenzi F, Freddi D (2024) Revelio: Interpretable long-form question answering. In: The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024. OpenReview.net, ???. https://openreview.net/forum?id=fyvEJXsaQf

Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: A human generated machine reading comprehension dataset. In: Besold TR, Bordes A, Garcez AS, Wayne G (eds) Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic

Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016. CEUR Workshop Proceedings, vol. 1773. CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

Niklaus J, Matoshi V, Rani P, Galassi A, Stürmer M, Chalkidis I (2023) LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pp 3016–3054. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.200

Nogueira R, Cho K (2019) Passage re-ranking with bert. arXiv preprint arXiv:1901.04085

Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D (eds) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. https://doi.org/10.3115/1073083.1073135 . https://aclanthology.org/P02-1040

Parida S, Motlícek P (2019) Abstract text summarization: A low resource challenge. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, pp 5993–5997. Association for Computational Linguistics. https://doi.org/10.18653/V1/D19-1616

Paszke A, Gross S, Massa F, Lerer A et al (2019) Pytorch: An imperative style, high-performance deep learning library. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) NeurIPS, pp 8024–8035

Puri R, Spring R, Patwary M, Shoeybi M, Catanzaro B (2020) Training question answering models from synthetic data. arXiv preprint arXiv:2002.09599

Ragazzi L, Italiani P, Moro G, Panni M (2024) What are you token about? differentiable perturbed top-k token selection for scientific document summarization. In: Findings of the Association for Computational Linguistics ACL 2024, pp 9427–9440

Ragazzi L, Moro G, Guidi S, Frisoni G (2024) Lawsuit: a large expert-written summarization dataset of italian constitutional court verdicts. Artif Intell Law

Ravichander A, Black AW, Wilson S, Norton TB, Sadeh NM (2019) Question answering for privacy policies: Combining computational and legal perspectives. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, pp 4946–4957. Association for Computational Linguistics. https://doi.org/10.18653/V1/D19-1500

Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M et al (1995) Okapi at trec-3. Nist Special Publication Sp 109:109

Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR arxiv:1910.01108

Schick T, Schütze H (2021) Generating datasets with pretrained language models. arXiv preprint arXiv:2104.07540

Shen Z, Lo K, Yu L, Dahlberg N, Schlanger M, Downey D (2022) Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. Adv Neural Inf Process Syst 35:13158–13173

Song K, Tan X, Qin T, Lu J, Liu T (2020) Mpnet: Masked and permuted pre-training for language understanding. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual

Sovrano F, Palmirani M, Distefano B, Sapienza S, Vitali F (2021) A dataset for evaluating legal question answering on private international law. In: Maranhão J, Wyner AZ (eds) ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, pp 230–234. ACM. https://doi.org/10.1145/3462757.3466094

Thomas P, Spielman S, Craswell N, Mitra B (2024) Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1930–1940

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Canton-Ferrer C, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T (2023) Llama 2: Open foundation and fine-tuned chat models. CoRR arxiv:2307.09288. https://doi.org/10.48550/ARXIV.2307.09288

Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, Huang S, Werra L, Fourrier C, Habib N et al (2023) Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944

Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M (2020) Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS

Wolf T, Debut L, Sanh V, Chaumond J et al (2020) Transformers: State-of-the-art natural language processing. In: EMNLP, pp 38–45. ACL, Online. https://doi.org/10.18653/v1/2020.emnlp-demos.6 . https://aclanthology.org/2020.emnlp-demos.6

Ye J, Gao J, Li Q, Xu H, Feng J, Wu Z, Yu T, Kong L (2022) Zerogen: Efficient zero-shot learning via dataset generation. arXiv preprint arXiv:2202.07922

Ye J, Gao J, Li Q, Xu H, Feng J, Wu Z, Yu T, Kong L (2022) Zerogen: Efficient zero-shot learning via dataset generation. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, pp 11653–11669. Association for Computational Linguistics. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.801

Yoo KM, Park D, Kang J, Lee S-W, Park W (2021) Gpt3mix: Leveraging large-scale language models for text augmentation. arXiv preprint arXiv:2104.08826

Yuan W, Neubig G, Liu P (2021) Bartscore: Evaluating generated text as text generation. In: Ranzato M, Beygelzimer A, Dauphin YN, Liang P, Vaughan JW (eds) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual, pp. 27263–27277

Yu Y, Zhuang Y, Zhang J, Meng Y, Ratner A, Krishna R, Shen J, Zhang C (2023) Large language model as attributed training data generator: A tale of diversity and bias. arXiv preprint arXiv:2306.15895

Zhang T, Kishore* V, Wu* F, Weinberger KQ, Artzi Y (2020) Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations. https://openreview.net/forum?id=SkeHuCVFDr

Zhang W, Shen H, Lei T, Wang Q, Peng D, Wang X (2023) GLQA: A generation-based method for legal question answering. In: International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, pp 1–8. IEEE. https://doi.org/10.1109/IJCNN54540.2023.10191483

Zheng L, Guha N, Anderson BR, Henderson P, Ho DE (2021) When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings. In: Maranhão J, Wyner AZ (eds) ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, pp 159–168. ACM. https://doi.org/10.1145/3462757.3466088

Zhou W, Zhang S, Gu Y, Chen M, Poon H (2023) Universalner: Targeted distillation from large language models for open named entity recognition. CoRR arxiv:2308.03279. https://doi.org/10.48550/ARXIV.2308.03279

Zhu B, Frick E, Wu T, Zhu H, Jiao J (2023) Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIF