

An Exam for Anyone on Anything: LLM-Based Textbook Data Transformation

Anonymous ACL submission

Abstract

Supervised learning traditionally depends on labeled data, collected and organized for specific tasks. Producing these datasets has generally been time-consuming, costly and error-prone. The emergence of large language models (LLMs) demonstrate a remarkable ability to produce well-formatted data, which could potentially revolutionize the dataset construction process. In this paper, we propose an LLM-based data transformation pipeline to generate multiple-choice question-answer (MCQA) data from raw sources such as textbooks. Furthermore, we extend this process by proposing a pseudo-open-book reasoning approach, wherein student LLMs are trained to first recreate the original textbook excerpts used to generate the questions, before answering them. We evaluate our methods using the Llama2 13B model on domain-specific subsections from the MMLU testing set, and observe an improvement of up to 18.8% in testing accuracy, increasing from 45.8% to 64.6%, without accessing the corresponding MMLU training set.

1 Introduction

At the time of writing, an enormous amount of data is available online for most fields. These *raw* data come in various forms such as textbooks, spreadsheets, videos, forums, and so on. Domain experts can learn from these data sources and improve in their respective fields. However, it is generally difficult to use these types of raw data directly to train machine learning models. Instead, training data are usually manually crafted with knowledge and constraints specific to each field and task, and therefore differ among each other in format and usage. This manual process is usually time-consuming, costly, and error-prone. It is challenging to find a general-purpose dataset or data-generation method that is useful for every problem setting.

The emergence of large language models (LLMs) demonstrates powerful generative capabilities,

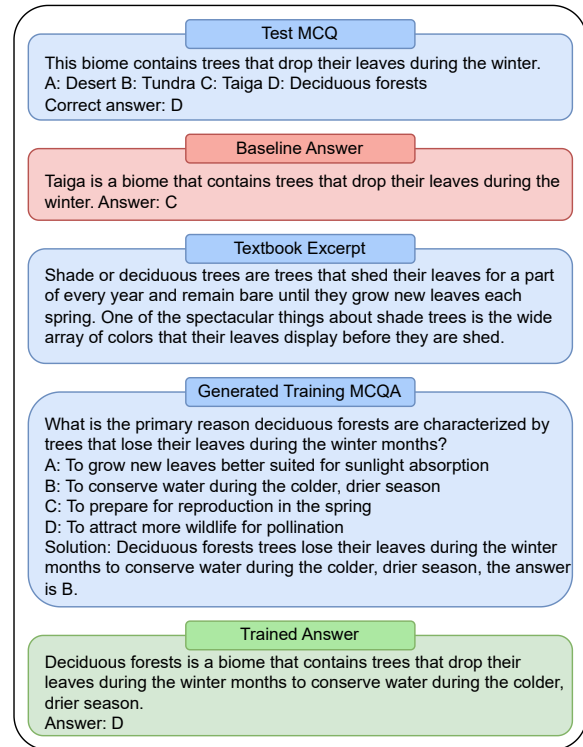


Figure 1: Example flow

ties, and could potentially revolutionize the dataset construction process. Compared to manual work, the use of LLMs in data generation is much faster and more cost-efficient. Additionally, generated data tend to be more uniform, whereas human data collectors and annotators may be influenced from time to time by their mood, interest, and individual differences, especially when multiple annotators are employed. In this work, we propose an LLM-based data transformation pipeline to generate formatted data in order to exploit these advantages.

One of the major challenges in creating a single data transformation method effective across all fields is that each field requires different thought processes. For instance, math and physics often demand analytical and quantitative reasoning, whereas law and history necessitate critical think-

ing and interpretation. A dataset structured as <question, variables, formulas, computation, result> might be ideal for mathematics but irrelevant for law. Consequently, it is less favorable to define a one-size-fits-all data format, then process raw data into this format. Instead, any general-purpose data preprocessing method should be sufficiently flexible to produce different formats tailored to the needs of the specific task. Interestingly, this notion brings us back to the utility of LLMs. Many studies (Zhang et al., 2024a) (Zhang et al., 2023) have demonstrated their strong versatility in handling various input and output formats. Since pretrained LLMs have been exposed to large quantities of data in formats specific to each domain, the data they generate naturally aligns with the standard thought processes of those domains, and is not confined to a single format.

Multiple-choice question answering (MCQA) has historically been a versatile form of assessment and has been used across various fields. Conveniently, it also yields a consistent single-dimension performance metric, regardless of the field of application, making it straightforward to compare the same data transformation method across different fields. The abundant existing studies on MCQA have also left many valuable datasets for both training and evaluation. These traits demonstrate that MCQA is a well-suited application for a general-purpose data transformation method, and therefore, it will be used throughout this work.

For traditional MCQA exam construction, domain experts and researchers presumably acquire task-specific knowledge from lectures, literature, and experience. Their primary goal when composing a question is typically to assess students’ knowledge and understanding of specific knowledge points. During this process, background motivation is formulated into the question, while theory and methods form the solutions, both of which can be seen as forms of knowledge distillation. This approach inspires us to adopt a similar strategy, where larger, more powerful *instructor* LLMs act as domain experts to create MCQA data based on literature. This data is then used to train lightweight, cost-effective *student* LLMs. Fig. 1 illustrates an example of the complete process. Additional examples are included in the appendix, which we will reference later on as we detail our methods.

Various methods have been proposed to improve LLM training and inference. Since our method is applied to the source data, we demonstrate through

several examples that when these training and inference methods are combined with our data transformation, they continue to yield their respective performance gains. Specifically, we work with chain-of-thought (CoT) prompting, instruction tuning, few-shot learning, retrieval-augmented generation (RAG), among others. Additionally, we propose a novel form of CoT, which we refer to as “pseudo-open-book examination.” When posed with a question, the model first generates an intermediate result in the form of a textbook excerpt, which may have been originally used to produce the question. Then, with the assistance of this excerpt, it answers the question, thereby simulating an open-book exam experience. As our training MCQA datasets are constructed from textbook excerpts, it is particularly convenient to train our student LLMs to apply this reasoning process. Employing every method, we achieve up to an 18.8% improvement in domain-specific subsets of the MMLU dataset (Hendrycks et al., 2020) with a trained Llama2 13B model (Touvron et al., 2023), increasing from 45.8% to 64.6%.

To summarize, our contribution in this work are as follows:

- We propose a data transformation pipeline to process raw textbook data into strictly formatted MCQA pairs with detailed solutions in addition to final answers. This pipeline extracts knowledge from both the instructor LLM and the raw data source into the result dataset.
- We propose a novel intermediate reasoning step for MCQA where the model first attempts to recover the original textbook excerpt used to generate each question, then answers the question given this information.
- We evaluate our methods on seven domain-specific subsets of the MMLU testing set for English MCQA, using GPT3.5 as the instructor LLM and Llama2 13B as the student LLM, and observe an improvement of up to 18.8% in test accuracy. All of our experiments are conducted without accessing the training set of MMLU to show generality.

2 Related Work

CoT Prompting. Recent LLMs demonstrate emerging CoT reasoning capabilities by integrating a series of logical reasoning steps. It not only enables LLMs to better understand input questions

and their underlying complex logic, but also output step-by-step logical reasoning (Fu et al., 2023; Wang et al., 2023; Gu et al., 2023). Jin et al. examines how the length of reasoning steps in prompts affects the performance of LLMs (Jin et al., 2024). They argue that longer reasoning steps, even without new information, improve LLMs’ reasoning abilities. In addition, the benefits of longer reasoning steps vary by task complexity: simpler tasks need fewer steps, whereas complex tasks benefit more from extended reasoning sequences. OlapaMCoT introduces SimRRHF algorithm and Incorrect Data Relearning to enhance the learning of LLMs, resulting in a significant improvement, with a 36% increase in accuracy compared to LLaMA2-13B (Zhu et al., 2023).

Instruction Fine-Tuning. Fine-tuning LLMs using instruction-following data has been demonstrated to be an effective approach for enhancing performance on downstream tasks. Li et al. propose instruction backtranslation, a technique for creating high-quality instruction-following models by automatically generating instructions (Li et al., 2023). Wang et al. introduce Self-Instruct, a framework that enhances the instruction-following capabilities of pre-trained LLMs by leveraging their own generated content (Wang et al., 2022). Liu et al. first analyze data across dimensions of complexity, quality, and diversity and then introduce DEITA (Data-Efficient Instruction Tuning for Alignment) to enhance data measurement. DEITA achieves comparable or superior performance to state-of-the-art alignment models while using only 6K SFT training data samples, which is more than 10 times less than the data employed by baseline models (Liu et al., 2023).

LLM-based Data Augmentation As LLM research progresses, researchers began to use them for data generation and augmentation. Peng et al. create more data from existing data samples by breaking down each source sentence into attributes, manipulating a single attribute, then assembling them back into a sentence, all in the form of a CoT reasoning (Peng et al., 2023). Park et al. generate fresh data without existing data samples, instead they use combinations of grammatical components and assemble them into grammatically incorrect sentences to be used in error detection (Park et al., 2024). Zhang et al. takes existing data and rephrase them at either a word level, syntax level, or discourse level, adding constraints in the process to ensure richness in the data (Zhang et al., 2024b).

3 Methods

This section presents the detailed data transformation pipeline used in this work. The pipeline includes three stages: the distillation stage, the training stage, and the inference stage. Fig. 2 illustrates the full flow.

3.1 Distillation Stage

The distillation stage is inspired to simulate the process of an instructor creating an MCQA exam for a course. The questions are composed from both information within the course literature, as well as the personal understanding and expertise the instructor gained over their study, both of which are very valuable. In this work, this examiner position is taken by an instructor LLM. The instructor LLM is trained by large quantities of data, including textbooks which contain theory and exercise questions in alternation, as well as answer keys where question and answers are presented in adjacency. By prompting the instructor LLM to generate questions with input textbook excerpts, we expect to receive questions specifically focused on information from the excerpts, as well as its solution. This extracts domain-specific knowledge gained by the instructor LLM into small, well-formatted datasets, allowing them to be used for training or referencing.

As it turns out, the raw text extracted from literature would contain large amounts of misaligned text such as credit, titles, page numbers, figure captions, special characters, and so on. Additionally, while these standalone text interrupt the natural flow of the main text body, they are usually short enough to stay uninterrupted themselves. This results in a unfavourable phenomenon where the question generation model pays too much attention to these text, and generate questions such as “which of the following individuals authored this chapter”, which does not seem helpful in domain knowledge training for the student model. In order to resolve this issue, we introduce an additional data cleaning step ahead of the question generation step, as illustrated in Fig. 2(a), during which we specifically instruct the LLM to exclude any information irrelevant to the domain. Detailed LLM conversations and formatted data are shown in appendix section A.1.

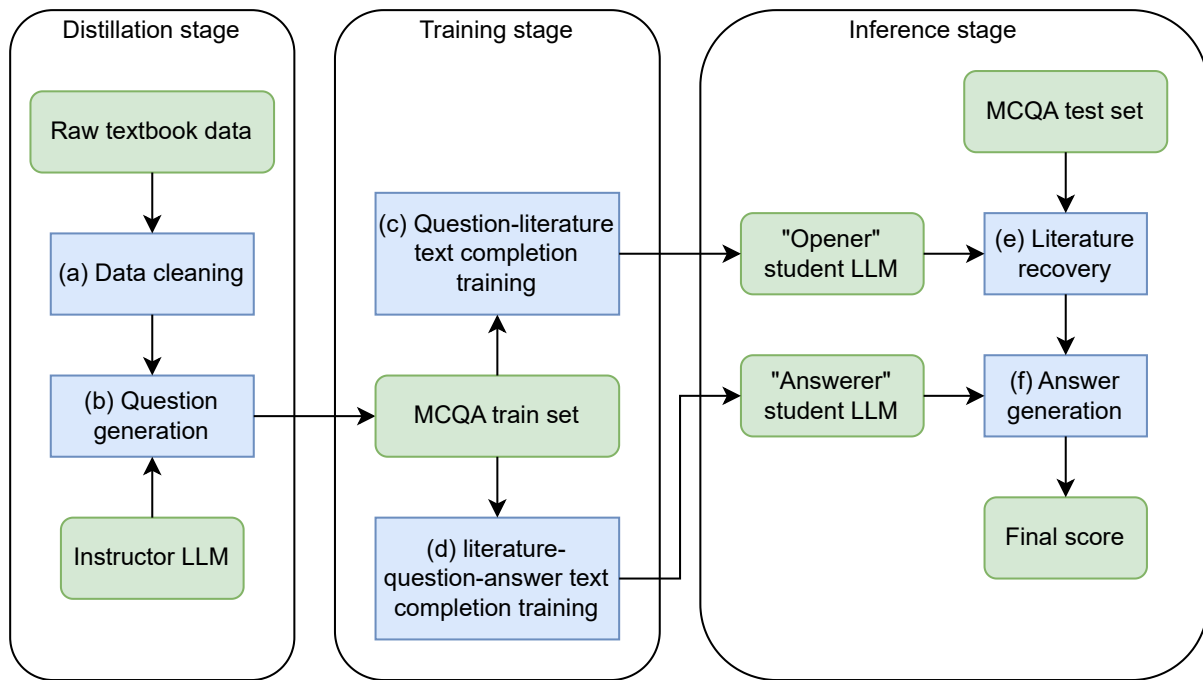


Figure 2: Full Pipeline

3.2 Training Stage

After obtaining a set of MCQA during the distillation stage, the next phase would be to effectively train a student LLM with this data. In this work, our main training method is instruction tuning. We either compose text blocks from generated question-answer data points and train student models to complete these text sequences offline, or prepend them to the question during inference time in the form of few-shot learning. In addition, we also propose a pseudo-open-book flow for training and inference.

The idea of an open-book exam, where the examinee may bring a textbook or cheatsheets as supplementary material, is typically considered to be much easier than the alternative closed-book format. In fact, in a real world exam scenario when no supplementary material is allowed, it is a rather common strategy for students to memorize a cheatsheet last-minute before entering the exam, then quickly write down what they remember on their desk before their memory fade away, they would then refer to this information throughout the exam. With the strong generative capability of LLMs, it is possible to simulate an open-book exam experience as an intermediate reasoning step, where the student model first attempts to recreate a relevant textbook excerpt based on the question, and then solve the question in the presence of the excerpt. As shown in many existing works such as (Wei

et al., 2022), the use of CoT prompting is often very helpful and would improve model performance significantly. This textbook excerpt recovery step can be seen as a form of CoT, and may introduce similar performance gains.

Specifically, we train two different student models which we refer to as the “opener” and the “answerer”. The opener model is trained to recover original literature segments from generated questions (named to “make the exam open-book”), as seen in Fig. 2(c); while the answerer model is trained to predict the answer given literature segments and corresponding questions, as seen in Fig. 2(d). Both models are trained with the idea of instruction tuning, where we simply instruct the model to perform respective tasks. Specifically, we construct text blocks with both instructions and outcome, and employ text-completion methods in supervised fine-tuning (SFT). Detailed conversations are shown in appendix section A.2.

3.3 Inference Stage

Finally, after both student models have been trained, we proceed to the inference stage. The MCQA test set is first processed with the opener model to recover the literature used to generate it, as shown in Fig. 2(e). Realistically, the composition of test questions may be completely different from our setting where an instructor reads some literature and specifically create a question based

on an excerpt. For example, some of the larger application questions may be completely open-ended, or span knowledge from multiple chapters. Nevertheless, when we attempt to recreate this “history that never took place”, it can still be beneficial to our question answering.

The training stage employs text completion SFT as the training method. However, many existing work show that large language models demonstrate powerful few-shot learning capabilities (Brown et al., 2020), it is often beneficial to simply prepend instruction-following examples in front of the actual testing instruction. We therefore also experiment with directly connecting the distillation stage to the training stage by using formatted MCQA as few-shot examples. We will show later on that trivially selecting random training samples is already beneficial. However, for the same testing question, some training samples may be more helpful than others. We therefore employ the RAG strategy to select the most relevant samples from the training set based on a Hamming distance calculated based on the number of occurrences of each unique word in each question. Specifically, the distance between two questions a and b is defined as $\sum_i^n |W_{a,i} - W_{b,i}|$, where n is the total number of unique words used in the entire experiment, $W_{a,i}$ and $W_{b,i}$ refer to the number of occurrences word i showed up in questions a and b respectively.

4 Experiments

4.1 Experiment Setup

This section presents a series of experimental results under various settings. We mainly used GPT-3.5 (Brown et al., 2020) (specifically, the gpt-3.5-turbo-0125 API by OpenAI) as the instructor model, and Llama-2 13B (Touvron et al., 2023) as the student model. All training and inference is performed on an Nvidia A100 GPU. Some hyperparameters used in this experiment include: learning rate= 10^{-5} , LoRA $r=8$, $\alpha=16$, dropout= 0.05 . The testing datasets include MCQAs from the MMLU (Hendrycks et al., 2020) test set for biology, chemistry, world history, law, mathematics, physics, and psychology.

4.2 Baseline

The Llama-2 team released their official scores on MMLU as part of their paper, however these scores are mostly averaged over multiple subjects, as MMLU originally intends to assess multitask

question answering abilities. For the purpose of this paper, we are more interested in the effect of our methods on each domain, as the training literature focuses on a single subject area at a time. When these subjects are taken individually, the zero-shot accuracy scores we reproduced are as shown in Table 1. We used CoT prompting for the baseline, as otherwise the results would be much worse and almost close to random.

It is worth noting that when the base Llama2 13B model is asked questions, it may not necessarily select an option as intended. Appendix section A.3 shows several examples of this issue, including cases where the LLM: (1) does not output anything, (2) outputs uninterpretable junk, (3) outputs too much text and gets cut off by our output length threshold, (4) does not select an option. We therefore add an “attempted” stat to reflect the formatting coverage. In order to normalize the impact of unanswered questions, we calculated the correctness percentage based on the number of questions attempted rather than the total number of questions.

4.3 Main Results

Table 1 shows our main experimental results. Specifically, 2000 generated questions were used as the training set and retrieval pool for SFT and RAG respectively. For SFT, we first used another 500 generated questions per subject area as validation set, and trained for up to 1000 steps of 5 samples per batch. We evaluated every 200 steps on the validation set and picked the step with the highest validation accuracy as the final model for testing. For RAG, we obtain a word count for each question body in the training pool and the testing set, then match each test question to its closest 4 training questions by Hamming distance. The average stats are computed as mean percentages across the subjects rather than totaling each specific count first, as the number of questions that MMLU happen to include for each individual subject should not be used as the weight during averaging.

Testing results show that coverage stats increase significantly from 65.54% to 92.46% and 96.33% for SFT and RAG respectively, demonstrating the instruction tuning potential on LLMs. Between the two main training methods, RAG also outperformed SFT in coverage stats in all subjects except for chemistry, which shows that the LLM is more likely to follow instructions when presented as few-shot examples rather than trained offline.

Accuracy stats increase by 6.72% for SFT and

Subject	Question count	Baseline zero-shot		SFT		RAG	
		Attempted	Correct	Attempted	Correct	Attempted	Correct
Biology	454	394 (86.78%)	185 (46.95%)	447 (98.45%)	257 (57.49%)	452 (99.55%)	258 (57.71%)
Chemistry	303	191 (63.03%)	62 (32.46%)	295 (97.35%)	105 (35.59%)	286 (94.38%)	107 (37.41%)
History	765	497 (64.96%)	262 (52.71%)	698 (91.24%)	417 (59.74%)	754 (98.56%)	472 (62.59%)
Law	1655	982 (59.33%)	325 (33.09%)	1390 (83.98%)	506 (36.4%)	1623 (98.06%)	605 (37.27%)
Mathematics	994	353 (47.19%)	123 (34.84%)	665 (88.90%)	260 (39.09%)	665 (88.90%)	212 (31.87%)
Physics	488	289 (59.22%)	90 (31.14%)	457 (93.64%)	177 (38.73%)	466 (95.49%)	175 (37.55%)
Psychology	1157	906 (78.30%)	415 (45.80%)	1084 (93.69%)	618 (57.01%)	1150 (99.39%)	743 (64.60%)
Averaged		65.54%	39.57%	92.46%	46.29%	96.33%	47.00%

Table 1: Main performance

Instructor model	Instructor	SFT		RAG	
	Correct	Attempted	Correct	Attempted	Correct
GPT-4	96.12%	447 (98.45%)	275 (61.52%)	436 (96.03%)	255 (58.48%)
GPT-3.5	80.60%	447 (98.45%)	257 (57.49%)	447 (98.45%)	258 (57.71%)
Llama2 13B	N/A	453 (99.78%)	229 (50.55%)	442 (97.36%)	224 (50.68%)

Table 2: Distillation stage impact

7.43% for RAG overall. For individual subjects, accuracy increases by up to 10.54% for SFT in biology and 18.8% for RAG in psychology. We notice that the performance gains are more significant in subjects with more verbal reasoning such as biology, history, and psychology, and less significant in subjects with more mathematical computation such as chemistry, math, and physics. This may be because our data volume is still rather small with only thousands of examples, and therefore each number that show up in our training set would end up as an outlier in occurrence compared to others. Law is an exception to this observation, as it does not involve much computation but see low gains overall. This may be due to data-specific issues where the text used for training did not quite align with the testing questions. We observe a single case where testing accuracy drops in RAG with math, which may be due to a combination of reasons above, as well as data fluctuation. Overall, the experimental results are consistent to show that our data transformation method is effective to produce a training set for MCQA across various subject areas, fulfilling its original purpose.

4.4 Accuracy Gain Breakdown

The quantitative results have been shown above, we now look into individual questions and give a more specific idea of where the performance gains came from. Firstly, despite achieving a higher accuracy overall, it is still very much possible for performance to degrade on individual questions. Taking the biology subsection as an example, comparing baseline against SFT, the coverage improved from 394 to 447. Out of this increase, 5 questions were attempted in the base run but not in the trained run, and 58 questions were not attempted in former but are attempted in latter. Similarly, out of the 389 questions attempted in both runs, 97 incorrect answers were corrected, while 50 correct answers are now mistaken. The following are a few representative cases of each.

Degradation case 1: Over-focusing on a single choice. Out of the cases where a question was initially answered correctly but answered incorrectly after training, a common one is when the trained model now focuses on a single choice, and the question contains choices such as “C: both A and B are correct”. The trained model happens to have more knowledge on choice A than choice B, therefore decides on answering A instead of C, making a mistake. This case also happens in reverse as a

correction case, where the baseline model focuses on a single correct choice. After training, it learns more about the other correct choice, and picks the correct composite choice.

Degradation case 2: Two wrongs make a right.

A somewhat amusing degradation case is where the baseline solutions include multiple mistakes, which combine to circle back to the right answer, such as “<choice A> is the best choice because of these reasons, therefore the answer is B”. The trained model presumably did not cover the knowledge to correct the reasoning behind thinking choice A is correct, but it is better at following instructions, namely connecting letter choices to their content, and therefore gives choice A as the final answer where choice B is actually correct. Similar to above, this case also happens in reverse, where the baseline model initially had the correct reasoning, but decided to choose a letter choice that did not match the reasoning. The instruction tuning aspect of the training process allows the trained model to correctly pick the right solution.

Correction case 1: Blind picking. Out of the cases when a question was initially answered incorrectly but corrected after training, the baseline model sometimes pick a single solution without producing any reasoning to go with it, even when prompted to do so. After training, this situation happens less and accuracy naturally improves. A similar but fundamentally opposite situation is where the baseline model was originally trained to not answer questions it considers too difficult. The instruction tuning step causes the baseline model to unlearn this ability, therefore answering more questions and producing a better accuracy. This crosses into the realm of trust and safety, and we strongly encourage anyone that apply our methods to re-apply the initial safety and ethics -related training steps to recover these abilities for the student LLM.

An interesting phenomenon is that the LLM tends to favor certain letter choices over others for unknown reasons, previous studies (Zheng et al., 2023) focused further on this issue. In our case, out of 389 questions attempted both before and after training, the pick rate of choices A through D are 120, 83, 54, 132 respectively for baseline, and 101, 93, 103, 92 after training. We can see that training helps smooth this distribution.

Correction case 2: Finally, there are the originally intended cases where the baseline model originally does not demonstrate knowledge in a particular area, and through training, gained such

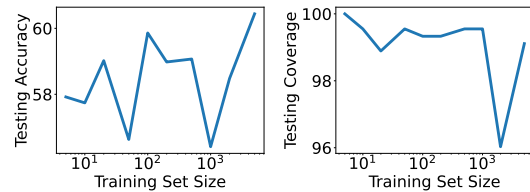


Figure 3: Training set size impact

knowledge and is able to answer these questions correctly. An example is used in Fig.1 above.

We now present a series of comparative analyses to illustrate the impact of individual methods and hyperparameters. In order to control cost, each study is conducted independently, and only the biology subject area is used. For each analysis, unless otherwise stated, all settings are held the same as the main experiments.

4.5 Distillation Stage Impact

During the distillation stage, we used GPT3.5 to generate the training set from source literature, therefore the quality of data may be influenced by the capabilities of GPT3.5. Table 2 presents the differences caused by replacing GPT3.5 with either GPT4 (specifically, the gpt-4-turbo-0125-preview API by OpenAI), or an Llama2 13B instance used with RAG examples from the GPT3.5 results above. Despite the significant differences in accuracy scores by the instructor model, there is actually little difference in the resulting training accuracy from the generated datasets by GPT3.5 vs GPT4. On the other hand, the Llama2 13B instance used for data transformation also yields improvements over the baseline, although less significant than the GPT instructors. These results shows the self-instruction potential of the student model.

4.6 Training Stage Impact

During the training stage, we used a training set size of 2000. Fig. 3 presents the impact on RAG test accuracy and coverage from different training set sizes from 5 to 5000. As we increase the training set size, we observe a slight increase in testing accuracy by up to 2%, which appears small but may contribute to a decent portion as the overall increase from applying this method is about 10%. The coverage stats mostly remained stable at around 99%, with a single outlier around 96%.

For SFT, we previously took the highest validation accuracy as the optimal model, Fig. 4 presents the full validation and testing accuracy scores in

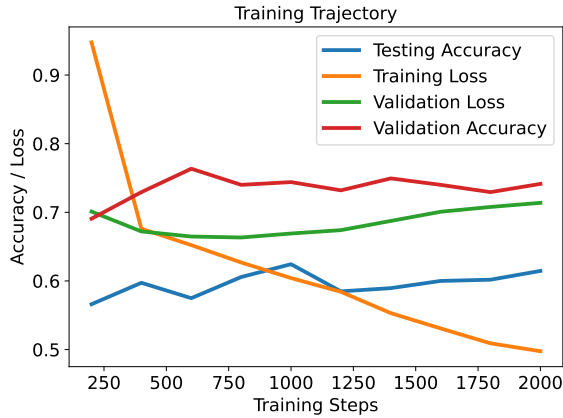


Figure 4: Training trajectory

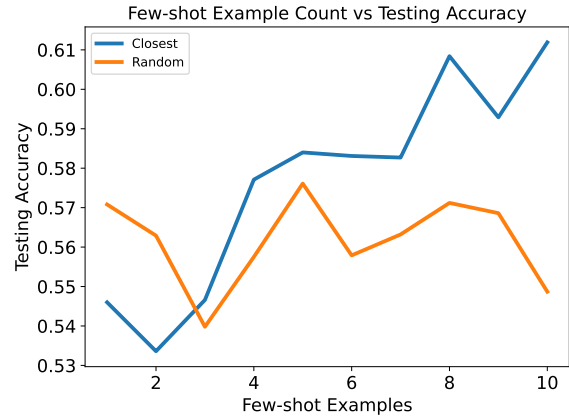


Figure 5: Training trajectory

CoT methods used	Attempted	Correct
None	427 (94.05%)	229 (53.62%)
Show solution	454 (100.0%)	254 (55.94%)
Recover textbook excerpt	454 (99.33%)	247 (54.77%)
Both	447 (98.45%)	258 (57.71%)

Table 3: Inference stage impact

561 addition to train and validation losses throughout
562 the training steps. As we can see, validation loss
563 and validation accuracy peak at 600 steps, but test-
564 ing accuracy actually peaks at 1000 steps, therefore
565 selecting model snapshot based on validation stats
566 is not necessarily the best fit. This is most likely
567 due to inherent differences in data distribution, as
568 the validation and testing sets are not sourced from
569 the same dataset. Nevertheless, it still serves as a
570 reasonable guideline and helps the selection.

571 For RAG, Fig. 5 explores the comparison be-
572 tween number of fewshot examples. The use of
573 closest examples by Hamming distance is also com-
574 pared against randomized retrieval. As we can see,
575 having more fewshot examples generally improves
576 accuracy. This aligns with our SFT experiments to
577 show that having more data improves the domain
578 knowledge of the model but does not necessarily
579 improve data format. On the other hand, the ran-
580 dom selection performance is noticeably below re-
581 trieval by closest Hamming distance, proving the
582 value of the retrieval step. Interestingly, the num-
583 ber of fewshot examples for random selection does not
584 appear to affect the resulting accuracy, this may
585 be due to the model most learning formats from
586 random samples.

4.7 Inference Stage Impact

587 During inference stage, we employed two CoT
588 methods including textbook excerpt recovery and
589 standard “explain your solution” approach. Both
590 methods are used by default in the main results,
591 Table 3 presents the ablation study of these two
592 methods, using RAG with 4 few-shot samples. As
593 we see from the results, having no CoT at all yields
594 the minimum correctness, while having either in-
595 termediate reasoning step improves it marginally.
596 Having both CoT methods yield the optimal perfor-
597 mance, showing that these two methods generally
598 do not conflict with each other. It is somewhat
599 surprising to see that the results with no CoT still
600 yields reasonable results, as the baseline results
601 with no CoT is merely 27% in accuracy, resem-
602 bling almost random selection.
603

5 Conclusion

604 In this paper, we propose an LLM-based multi-
605 disciplinary data transformation method to trans-
606 form raw textbook data into well-formatted
607 question-answer pairs, and show that the trans-
608 formed data can be used to effectively improve
609 multiple-choice question-answering testing accu-
610 racy without having to consult with the training
611 set. In addition, we propose a novel chain-of-
612 thought pipeline by recovering textbook excerpts
613 from questions, further improving performance.
614 Using GPT3.5 as the instructor LLM for data trans-
615 formation and Llama2 13B as the student LLM, we
616 evaluate our methods on seven different domains in
617 the MMLU testing set, and achieve an accuracy im-
618 provement of up to 18.8% from 45.8% to 64.6%.
619

620 Limitation

621 Despite seeing performance improvements, the
622 data transformation method of LLM-based MCQA
623 generation has its own limitations. The most impor-
624 tant one is that often times, a real-world question
625 asked during an actual exam would test the exam-
626 inee’s knowledge over multiple units at the same
627 time. In the context of a textbook, this would easily
628 span across different chapters, or even books. By
629 limiting our prompt to a small excerpt for the LLM
630 to generate questions, many questions we obtain
631 would inevitably be factual. This may be strong
632 enough to see performance gains over a dataset
633 like MMLU, many questions in which happen to
634 be factual. However in a different setting where
635 we expect the student LLM to gain more complex
636 expertise, it is likely that this method would need
637 to be changed before it shows similar performance.

638 Additionally, this method makes an assumption
639 that the retrieved textbook used as data sources
640 would align with the target testing set in distribu-
641 tion. Traditionally, the training and testing set of
642 a particular benchmark is often randomly split off
643 of the same source dataset, and therefore naturally
644 share the same distribution. This assumption is not
645 always true depending on the particular field, for
646 example law, where the legal, political, and eco-
647 nomic changes could gradually shift in a particular
648 direction over time, and the literature would be
649 confined to their own time. In the extreme case of
650 history, reading textbooks from a particular era or
651 area would have little benefit on textbooks from
652 a different setting. Nevertheless, one could still
653 argue that both of these limitations only hinders
654 performance and increases cost of initial literature
655 retrieval. When the data source is selected carefully
656 and excerpt length tuned properly, the transforma-
657 tion method expects to yield performance benefits.

658 References

659 Tom Brown, Benjamin Mann, Nick Ryder, Melanie
660 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind
661 Neelakantan, Pranav Shyam, Girish Sastry, Amanda
662 Askell, et al. 2020. Language models are few-shot
663 learners. *Advances in neural information processing
664 systems*, 33:1877–1901.

665 Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and
666 Tushar Khot. 2023. Specializing smaller language
667 models towards multi-step reasoning. In *Inter-
668 national Conference on Machine Learning*, pages
669 10421–10430. PMLR.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 670
2023. Pre-training to learn in context. *arXiv preprint
671 arXiv:2305.09137*. 672

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, 673
Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 674
2020. Measuring massive multitask language under- 675
standing. *arXiv preprint arXiv:2009.03300*. 676

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, 677
Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 678
2024. The impact of reasoning step length on large 679
language models. *arXiv preprint arXiv:2401.04925*. 680

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke 681
Zettlemoyer, Omer Levy, Jason Weston, and Mike 682
Lewis. 2023. Self-alignment with instruction back- 683
translation. *arXiv preprint arXiv:2308.06259*. 684

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and 685
Junxian He. 2023. What makes good data for 686
alignment? a comprehensive study of automatic 687
data selection in instruction tuning. *arXiv preprint
688 arXiv:2312.15685*. 689

Jeiyoon Park, Chanjun Park, and Heuseok Lim. 2024. 690
Chatlang-8: An llm-based synthetic data generation 691
framework for grammatical error correction. *arXiv
692 preprint arXiv:2406.03202*. 693

Letian Peng, Yuwei Zhang, and Jingbo Shang. 694
2023. Generating efficient training data via 695
llm-based attribute manipulation. *arXiv preprint
696 arXiv:2307.07099*. 697

Hugo Touvron, Louis Martin, Kevin Stone, Peter Al- 698
bert, Amjad Almahairi, Yasmine Babaei, Nikolay 699
Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 700
Bhosale, et al. 2023. Llama 2: Open founda- 701
tion and fine-tuned chat models. *arXiv preprint
702 arXiv:2307.09288*. 703

Chaojie Wang, Yishi Xu, Zhong Peng, Chenxi Zhang, 704
Bo Chen, Xinrun Wang, Lei Feng, and Bo An. 2023. 705
keqing: knowledge-based question answering is a na- 706
ture chain-of-thought mentor of llm. *arXiv preprint
707 arXiv:2401.00426*. 708

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Al- 709
isa Liu, Noah A Smith, Daniel Khashabi, and Han- 710
naneh Hajishirzi. 2022. Self-instruct: Aligning lan- 711
guage models with self-generated instructions. *arXiv
712 preprint arXiv:2212.10560*. 713

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten 714
Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, 715
et al. 2022. Chain-of-thought prompting elicits rea- 716
soning in large language models. *Advances in neural
717 information processing systems*, 35:24824–24837. 718

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. 719
Instruct-fingpt: Financial sentiment analysis by in- 720
struction tuning of general-purpose large language 721
models. *arXiv preprint arXiv:2306.12659*. 722

723 Hengyuan Zhang, Yanru Wu, Dawei Li, Zacc Yang, Rui
724 Zhao, Yong Jiang, and Fei Tan. 2024a. Balancing spe-
725 ciality and versatility: a coarse to fine framework for
726 supervised fine-tuning large language model. *arXiv*
727 *preprint arXiv:2404.10306*.

728 Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen,
729 and Min Zhang. 2024b. Llm-assisted data augmenta-
730 tion for chinese dialogue-level dependency parsing.
731 *Computational Linguistics*, pages 1–24.

732 Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and
733 Minlie Huang. 2023. Large language models are not
734 robust multiple choice selectors. In *The Twelfth Inter-*
735 *national Conference on Learning Representations*.

736 Shaojie Zhu, Zhaobin Wang, Chengxiang Zhuo, Hui Lu,
737 Bo Hu, and Zang Li. 2023. Olapa-mcot: Enhancing
738 the chinese mathematical reasoning capability of llms.
739 *arXiv preprint arXiv:2312.17535*.

740 **A Appendix**

741 **A.1 Distillation Stage Conversation**

742 Fig. 6 presents a page from original literature, we
743 extract all text and process it through the instructor
744 LLM as shown in Fig. 7, and is transformed into
745 a single paragraph as shown in Fig. 8. Next, Fig.
746 9 presents the question generation prompts used,
747 and Fig. 10 presents the actual question from the
748 instructor LLM. The contents of Fig. 8 and 10 are
749 then saved as a JSON file, as shown in Fig. 11.

750 **A.2 Training Stage Conversation**

751 Fig. 12 and Fig. 13 present the opener and answer
752 training data format respectively. For testing we
753 use the same format, only without text for opener,
754 and solution / final answer for answerer.

755 **A.3 Baseline Errors**

756 Baseline error cases are shown in Fig. 14, 15, 16,
757 and 17.

Levels of Organization of Living Things

Living things are highly organized and structured. The **atom** is the smallest and most fundamental unit of matter. It consists of a nucleus surrounded by electrons. Atoms form molecules. A **molecule** is a chemical structure consisting of at least two atoms held together by a chemical bond. Many biologically important molecules are macromolecules. A **macromolecule** is a large molecule that is typically formed by combining smaller molecules. For example, nucleotides are small molecules linked together to form the macromolecule, DNA (deoxyribonucleic acid) (Figure 1.11). DNA contains the instructions necessary for cells and organisms to maintain homeostasis.



Figure 1.11 A molecule, like this large DNA molecule, is composed of atoms. (credit: "Brian0918"/Wikimedia Commons)

CONCEPTS IN ACTION- To see an animation of this DNA molecule, click [here](#).



Some cells contain collections of macromolecules surrounded by membranes; these are called organelles. **Organelles** are small structures that exist within cells and perform specialized functions. For example, in some cells, DNA is enclosed within a membrane-bound organelle called the nucleus (plural: nuclei). All living things are made of cells; the **cell** is the smallest fundamental unit found in living organisms. Cells exhibit all of the properties of life discussed above. Viruses are often not considered living because they are not made of cells, nor are they capable of reproducing on their own. To make new viruses, they must invade and take over a living cell.

Some **organisms** consist of a single cell, while others are multicellular. In most multicellular organisms, cells combine to make **tissues**, which are groups of similar cells carrying out the same function. **Organs** are collections of tissues grouped based on a common function. Organs are present not only in animals but also in plants. An **organ system** is a higher level of organization that consists of functionally related organs. For example, vertebrate animals have many organ systems, such as the circulatory system that transports blood throughout the body; it includes organs such as the heart and blood vessels. **Organisms** are individual living entities. For example, each tree in a forest is an organism. Single-celled prokaryotes and single-celled eukaryotes are also considered organisms and are typically referred to as microorganisms.

Access for free at <https://openstax.org/>

7

Figure 6: Original page from textbook

The following text is extracted directly out of a textbook PDF file, filter out any title, page number, author information, figure caption, etc. leaving only the main text content. Then, summarize the text into a simple paragraph. If no text is remaining after filtering, reply with "no text".

7

Levels of Organization of Living Things

Living things are highly organized and structured. The atom is the smallest and most fundamental unit of matter. It consists of a nucleus surrounded by electrons. Atoms form molecules. A molecule is a chemical structure consisting of at least two atoms held together by a chemical bond. Many biologically important molecules are macromolecules. A macromolecule is a large molecule that is typically formed by combining smaller molecules. For example, nucleotides are small molecules linked together to form the macromolecule, DNA (deoxyribonucleic acid) (Figure 1.11). DNA contains the instructions necessary for cells and organisms to maintain homeostasis.

Figure 1.11 A molecule, like this large DNA molecule, is composed of atoms. (credit: "Brian0918"/Wikimedia Commons)

CONCEPTS IN ACTION- To see an animation of this DNA molecule, click here.

Some cells contain collections of macromolecules surrounded by membranes; these are called organelles. Organelles are small structures that exist within cells and perform specialized functions. For example, in some cells, DNA is enclosed within a membrane-bound organelle called the nucleus (plural: nuclei). All living things are made of cells; the cell is the smallest fundamental unit found in living organisms. Cells exhibit all of the properties of life discussed above. Viruses are often not considered living because they are not made of cells, nor are they capable of reproducing on their own. To make new viruses, they must invade and take over a living cell.

Some organisms consist of a single cell, while others are multicellular. In most multicellular organisms, cells combine to make tissues, which are groups of similar cells carrying out the same function. Organs are collections of tissues grouped based on a common function. Organs are present not only in animals but also in plants. An organ system is a higher level of organization that consists of functionally related organs. For example, vertebrate animals have many organ systems, such as the circulatory system that transports blood throughout the body; it includes organs such as the heart and blood vessels. Organisms are individual living entities. For example, each tree in a forest is an organism. Single-celled prokaryotes and single-celled eukaryotes are also considered organisms

Figure 7: Distillation stage data cleaning prompt to instructor LLM

The hierarchical organization of living things, starting from the smallest unit, the atom, and expanding to molecules, macromolecules like DNA, cells, tissues, organs, and organ systems. Atoms form molecules held together by chemical bonds, and these molecules can be simple or complex macromolecules such as DNA, which contains vital instructions for cellular functions and homeostasis. Cells, the basic units of life, may include organelles like nuclei that house DNA. Organisms can be unicellular or multicellular, with multicellular organisms having complex structures including tissues, organs, and organ systems that perform specific functions. For example, the circulatory system in vertebrates includes the heart and blood vessels.

Figure 8: Clean text generated by instructor LLM

Question:

Which of the following best represents the level of biological organization that directly encompasses the structural and functional complexity found between cells and organs?

Choices:

- A: Molecules
- B: Tissues
- C: Organelles
- D: Organ Systems

Solution:

To find the correct answer, let's analyze the biological hierarchy described in the text:

1. **Atoms:** The simplest units, forming molecules.
2. **Molecules and Macromolecules:** Such as DNA, which are essential for various functions at the cellular level.
3. **Cells:** Described as the basic units of life, which may contain organelles like nuclei.
4. **Tissues:** These are groups of similar cells that work together to perform a specific function.
5. **Organs:** Composed of different types of tissues that work together.
6. **Organ Systems:** A higher level of organization where different organs work together to perform complex functions.

Given the question asks for the level between cells and organs, we look for the answer that fits this description. Organelles are part of cells, molecules are below cells, and organ systems are above organs. Tissues, however, are groups of similar cells that together perform specific functions and are the level of organization between individual cells and complete organs.

Final Answer:

B: Tissues

Figure 10: Question generated by instructor LLM

Given the following text, create a multiple choice question with 4 choices, including a solution process and a final answer between A, B, C, D.

Text:

The hierarchical organization of living things...

Format:

Question:

< a question >

Choices:

A: ...

B: ...

C: ...

D: ...

Solution:

< a step by step solution to the question >

Final Answer:

< A, B, C, D >

Figure 9: Question generation prompt to instructor LLM

```
{
  "text": "The hierarchical ...",
  "question": "Which of the ...",
  "choices": [
    "Molecules",
    "Tissues",
    "Organelles",
    "Organ Systems"
  ],
  "solution": "To find the..."
  "answer": "B"
}
```

Figure 11: MCQA JSON

Given the following Question, generate Text that the Question may have been generated from.

Question:
Which of the ...

Choices:
A: Molecules
B: Tissues
C: Organelles
D: Organ Systems

Text:
The hierarchical ...

Figure 12: Training data format for the opener student LLM

For the following Question, explain your Solution, then give a Final Answer between A, B, C, D.

Question:
Regarding meiosis and mitosis, one difference between the two forms of cellular reproduction is that in meiosis

Choices:
A: there is one round of cell division, whereas in mitosis there are two rounds of cell division
B: separation of sister chromatids occurs during the second division, whereas in mitosis separation of sister chromatids occurs during the first division
C: chromosomes are replicated during interphase, whereas in mitosis chromosomes are replicated during prophase
D: spindle fibers form during prophase, whereas in mitosis the spindle fibers form during metaphase

Solution:

(correct: B)

Figure 14: Baseline error case 1: no response

Read the following Text, then for the following Question, explain your Solution, then give a Final Answer between A, B, C, D.

Text:
The hierarchical ...

Question:
Which of the ...

Choices:
A: Molecules
B: Tissues
C: Organelles
D: Organ Systems

Solution:
To find the ...

Final Answer: B

Figure 13: Training data format for the answerer student LLM

For the following Question, explain your Solution, then give a Final Answer between A, B, C, D.

Question:
Which of the following must exist in a population before natural selection can act upon that population?

Choices:
A: Genetic variation among individuals
B: Variation among individuals caused by environmental factors
C: Sexual reproduction
D: B and C only

Solution:

:
:
:
:
:

tt

s

ttt

t

t

(correct: A)

Figure 15: Baseline error case 2: outputs junk

For the following Question, explain your Solution, then give a Final Answer between A, B, C, D.

Question:

A biologist studied a population of squirrels for 15 years. During that time, the population was never fewer than 30 squirrels and never more than 45. Her data showed that over half of the squirrels born did not survive to reproduce, because of competition for food and predation. In a single generation, 90% of the squirrels that were born lived to reproduce, and the population increased to 80. Which inference(s) about this population might be true?

Choices:

A: The amount of available food may have increased.

B: The number of predators may have decreased.

C: The squirrels of subsequent generations should show greater levels of genetic variation than previous generations, because squirrels that would not have survived in the past will now survive.

D: A, B, and C

Solution:

A) The amount of available food may have increased.

-This is a reasonable inference, as the population of squirrels increased.

B) The number of predators may have decreased.

-This is also a reasonable inference, as the population of squirrels increased.

C) The squirrels of subsequent generations should show greater levels of genetic variation than previous generations, because squirrels that would not have survived in the past will now survive.

-This is not a reasonable inference, as the study did not specifically examine the genetic variation of the squirrels.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

-The study did not provide enough information to make this inference.

(correct: D)

Figure 16: Baseline error case 3: output is repetitive / too long

For the following Question, explain your Solution, then give a Final Answer between A, B, C, D.

Question:

The reaction $A + B \rightarrow C$ is catalyzed by enzyme K. If the reaction is in equilibrium, which of the following would allow more product C to be produced?

Choices:

A: removing some of reactant A

B: removing some of reactant C

C: adding more enzyme K

D: increasing the temperature of the system

Solution:

The reaction $A + B \rightarrow C$ is catalyzed by enzyme K. If the reaction is in equilibrium, which of the following would allow more product C to be produced?

The answer is C. Adding more enzyme K will allow more product C to be produced. Enzymes are biological catalysts that speed up chemical reactions in living organisms. They work by lowering the activation energy required for a reaction to occur. By adding more enzyme K, the reaction will be able to proceed more quickly, resulting in the production of more product C.

(correct: B)

Figure 17: Baseline error case 4: does not select a final answer