Agentic Adversarial QA for Improving Domain-Specific LLMs

Vincent Grari [1 2] **Ciprian Tomoiagă** [1] **Sylvain Lamprier** [3] **Tatsunori Hashimoto** [4] **Marcin Detyniecki** [1 5 2]

Abstract

Large Language Models (LLMs), despite extensive pretraining on broad internet corpora, often struggle to adapt effectively to specialized domains. There is growing interest in fine-tuning these models for such domains; however, progress is constrained by the scarcity and limited coverage of high-quality, task-relevant data. To address this, synthetic data generation methods such as paraphrasing or knowledge extraction are commonly applied. Although these approaches excel at factual recall and conceptual knowledge, they suffer from two critical shortcomings: (i) they provide minimal support for interpretive reasoning capabilities in these specialized domains, and (ii) they often produce synthetic corpora that are excessively large and redundant, resulting in poor sample efficiency. To overcome these gaps, we propose an adversarial question-generation framework that produces a compact set of semantically challenging questions. These questions are constructed by comparing the outputs of the model to be adapted and a robust expert model grounded in reference documents, using an iterative, feedback-driven process designed to reveal and address comprehension gaps. Evaluation on specialized subsets of the LegalBench corpus demonstrates that our method achieves greater accuracy with substantially fewer synthetic samples.

1. Introduction

Large Language Models (LLMs) pretrained on extensive internet-scale corpora have demonstrated remarkable general-purpose capabilities but often struggle to efficiently adapt to highly specialized, domain-specific knowledge contained within small, focused document sets. This difficulty arises primarily from the models' unfamiliarity with critical domain-specific facts—which may appear only sparsely or even just once within a limited corpus—making it especially challenging to interpret and integrate these facts through complex reasoning. Although there is growing interest in fine-tuning LLMs for specific domains, progress has been limited due to the scarcity of specialized data, motivating the development of targeted synthetic augmentation approaches.

Recent approaches, such as EntiGraph (Yang et al., 2024b) and Knowledge-Instruct (Ovadia et al., 2025), aim to close the data-efficiency gap by generating synthetic training data from small corpora, typically through structured entity and fact extraction. These methods commonly leverage a large, general-purpose model to generate new training data grounded in the specialized domain documents, which is then used to improve a smaller, domain-specialized model, often via fine-tuning or distillation. While EntiGraph focuses on expanding entity-centric knowledge graphs for continued pretraining, Knowledge-Instruct reformulates extracted knowledge into instruction-response pairs for supervised fine-tuning. Although these approaches achieve strong performance on tasks focused primarily on factual recall, their effectiveness may diminish when faced with broader comprehension tasks, especially those requiring nuanced interpretation, inference, and integration of complex domain-specific knowledge. For example, determining insurance coverage for an uncommon event, such as a client experiencing an accident while hiking abroad, requires interpreting rarely encountered or implicitly defined conditions within policy clauses. Resolving such questions involves not merely recalling explicit facts, but understanding nuanced language, subtle contextual implications, and layered dependencies within domain-specific documentation.

This observation suggests that it is not just the quantity of data that matters, but rather the quality and specificity of the augmentation strategies employed, particularly approaches explicitly designed to pinpoint and address specific model deficiencies. Existing augmentation methods generally generate synthetic data indiscriminately or based on pre-defined heuristics, without directly probing which concepts or reasoning tasks are most difficult for the model. This can lead to inefficient training, as much of the augmented data might reinforce knowledge the model already possesses. To address this, an adversarial approach naturally emerges as a promis-

[1]AXA Group Operations [2]TRAIL, Sorbonne Université, Paris, France [3]LERIA, Université d'Angers, France [4]Stanford University [5]Polish Academy of Science, IBS PAN, Warsaw, Poland. Correspondence to: Vincent Grari <vincent.grari@axa.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ing alternative: by actively generating challenging questions tailored to expose precisely those aspects of reasoning and comprehension where the model exhibits weaknesses. In this manner, the model is exposed not only to questions that require factual recall but also, and more importantly, to those that demand interpretive and integrative reasoning. Inspired by principles from active learning (Xu et al., 2013), boosting (Freund & Schapire, 1997), and distributionally robust optimization (DRO) (Duchi & Namkoong, 2018; Sinha et al., 2018)—which enhance learning by focusing on uncertain, misclassified, or worst-case instances where the model may perform poorly—we introduce an adversarial learning framework for smaller domain-specific LLMs. This framework iteratively generates questions, using feedback from a robust expert model, to systematically uncover and address interpretative weaknesses.

We demonstrate our adversarial feedback-driven methodology on specialized legal documents within the LegalBench corpus (Guha et al., 2023), showing how these targeted synthetic datasets substantially improve the domain-specific reasoning capabilities of smaller LLMs. Our method enables these models to achieve performance competitive with much larger counterparts, surpassing entity-centric and instruction-based strategies by more effectively addressing a wider range of nuanced comprehension challenges.

2. Related Work

Transfer Learning Across Domains. A core challenge in natural language processing is adapting language models to new domains where direct supervision or labeled data is scarce. Traditional approaches for cross-domain transfer learning often involve continued pretraining on domain corpora (Gururangan et al., 2020; Devlin et al., 2019), domain-adaptive fine-tuning (Howard & Ruder, 2018; Lee et al., 2020), or leveraging multi-task learning frameworks (Ruder, 2017). Such methods aim to bridge the gap between general pretraining and the specific terminology, knowledge, or reasoning patterns required in target domains. However, their effectiveness may be constrained by the limited availability of high-quality domain-specific data, motivating the development of more data-efficient adaptation strategies.

Model Distillation and Student-Teacher Paradigms. To further enhance transfer across domains, model distillation methods have been widely employed (Lin et al., 2021; Sanh et al., 2019). In this setup, a large, general-purpose teacher model transfers knowledge to a smaller, specialized student model, typically by having the student mimic the outputs or intermediary representations of the teacher. Recent work has explored domain-adaptive distillation (Turc et al., 2019; Jiao et al., 2020), where the teacher is either fine-tuned or prompted for domain-specific tasks, and the resulting guidance is used to supervise smaller models with reduced resources. These teacher-student setups are related to our approach, but in our case, the teacher model is grounded in the same domain documents as the student. This ensures the feedback is tailored to the specific context, making the supervision more relevant and effective.

Synthetic Data Generation. Recent approaches adapt language models to specialized domains with limited data by generating synthetic corpora. Entity-centric methods such as EntiGraph (Yang et al., 2024b) expand on entities and relations via knowledge graphs, while paraphrase-based techniques (Maini et al., 2024; Ovadia et al., 2024) diversify training data by rewording existing texts. Instruction-driven approaches like Knowledge-Instruct (Ovadia et al., 2025) convert extracted domain facts into instruction–response pairs, supporting efficient adaptation under low-resource settings. By contrast, our adversarial, feedback-driven framework improves sample efficiency and reasoning ability by selectively generating questions that require not only factual recall but also deeper context-dependent reasoning.

Alongside these developments, there is increasing interest in using LLMs in iterative feedback loops to refine prompts and tasks. Such feedback-driven frameworks, where the model provides critiques or guidance, underlie our approach.

Prompt Optimization via LLM Feedback. Recent approaches optimize prompts and inputs using automated feedback generated by large language models. In these frameworks, prompt refinement are guided by structured critiques or optimization signals from external LLMs. For instance, methods such as TextGrad (Yüksekgönül et al., 2024) implement differentiable prompting, iteratively refining queries to maximize a model-evaluated reward. Other techniques leverage LLM feedback loops for instruction tuning and robust evaluation (Yang et al., 2024a; Madaan et al., 2023). Our adversarial question generation pipeline builds upon this line of work, using LLM-driven feedback to provide fine-grained, diagnostic supervision in an adversarial optimization framework.

3. Methodology

Our objective is to enhance both the *domain-specific knowledge*—the understanding and recall of concepts, facts, and language unique to a given specialized corpus—and the *interpretive reasoning capabilities*—that is, the ability to accurately integrate, infer, and reason over such content—of language models, based solely on access to a domain document C (e.g., a legal contract). As in prior work, we operate solely with the domain corpus itself, without requiring annotated datasets or downstream task supervision.

To systematically address both types of weaknesses, our approach introduces an iterative adversarial process that generates questions $Q^{(t)}$ targeting the model's factual gaps

and interpretative limitations. For example, some questions may focus on domain-specific terminology, concepts, or facts explicitly present in C (e.g., "What is the definition of the term 'subrogation' in this contract?"), while others may present challenging or hypothetical scenarios that require reasoning, inference, or integration of multiple clauses (e.g., "If a client is injured while hiking abroad, would they be covered under this insurance policy?"). A model may, for instance, accurately define 'subrogation' but fail when asked to determine coverage in a scenario requiring integration of multiple clauses; an adversarial question targeting this scenario reveals such interpretive limitations. In this context, the effectiveness of adversarial question generation may be conceptually likened to pedagogical practices in active learning: rather than solely requiring the passive recall of explicitly stated definitions and isolated facts—analogous to students memorizing material for traditional written exams—comprehensive understanding is better gauged by posing challenging, integrated, or hypothetical scenarios akin to those one might encounter in rigorous oral examinations.

In our setting, we assume access to a large, highly capable language model which, when provided with the full domain corpus C as context, can function as a robust "oracle" or expert. However, relying on such a system—in which the model always has access to all domain documents—is often impractical in real-world scenarios due to resource, latency, privacy, or deployment constraints. Therefore, our aim is to transfer domain expertise from this strong oracle model to a more lightweight, efficient target model.

3.1. Adversarial Question Optimization

We propose an adversarial optimization framework to systematically uncover and address interpretive deficiencies in domain-specific language models. This setup comprises two primary agents: a robust expert model (f_{strong}) and a target weaker model (f_{weak}), both provided with access to the same domain-specific context C. The core objective is to generate and iteratively refine questions about C that maximize divergence between the responses of these models, thereby identifying aspects in which the target model exhibits limitations in either domain-specific knowledge or interpretive reasoning. Formally, at each iteration t, we take the domain-specific context or document C and the current question $Q_i^{(t)}$, and obtain answers from both models:

$$A_{i,\text{strong}}^{(t)} = f_{\text{strong}}(C, Q_i^{(t)}), \quad A_{i,\text{weak}}^{(t)} = f_{\text{weak}}(C, Q_i^{(t)}).$$

We then evaluate the difference in their answers using a feedback function f_{fb}, which is typically instantiated as a capable LLM to compare the two responses and identify discrepancies along dimensions such as correctness, coverage, and contextual reasoning alignment (see Appendix A.3.3):

$$\mathcal{L}(Q_i^{(t)}) = f_{\text{fb}}(A_{i,\text{strong}}^{(t)}, A_{i,\text{weak}}^{(t)}).$$

The key objective of this adversarial process is to generate questions that maximize the measured disagreement, as quantified by $\mathcal{L}(Q)$, which serves as a proxy loss function indicating where the target model most strongly diverges from the expert's interpretive capacity. Specifically, we seek

$$Q^* = \arg\max_Q \mathcal{L}(Q).$$

To operationalize this iterative maximization over text, we adopt the differentiable prompting paradigm introduced in TextGrad (Yüksekgönül et al., 2024). Notably, while TextGrad is designed to minimize a task loss by refining prompts, our framework inverts this direction and explicitly maximizes interpretive disagreement in order to expose the limitations of the target model:

$$Q_i^{(t+1)} = Q_i^{(t)} + \nabla_Q \mathcal{L}(Q_i^{(t)}).$$

Each refinement step comprises two stages: (i) a guidance model (f_{guide}) generates a natural language editing instruction, conditioned on the output of the feedback model (f_{fb}): $\nabla_Q \mathcal{L}(Q_i^{(t)}) = f_{\text{guide}}(Q_i^{(t)}, \mathcal{L}(Q_i^{(t)}))$. This instruction prescribes how to revise $Q_i^{(t)}$ to accentuate potential weaknesses and failure modes in the target model (see Appendix A.3.3 for examples). (ii) a revision model (f_{rev}) applies this instruction to update the current question, yielding the next iteration, $Q_i^{(t+1)} = f_{\text{rev}}(Q_i^{(t)}, \nabla_Q \mathcal{L}(Q_i^{(t)}))$.

In practice, we instantiate the three auxiliary agents—f_{fb}, f_{guide}, and f_{rev}—using the same strong LLM, each configured with a distinct instruction prompt (see Appendix A.3.3). This design preserves semantic consistency across modules while remaining implementation-lightweight. Notably, f_{guide} and f_{rev} correspond to the `TextGrad.backward` and `TextGrad.step` modules, respectively.

Algorithm 1 summarizes the entire adversarial question optimization procedure, clearly delineating each step involved in refining questions to systematically identify and improve the interpretive limitations of the target model.

3.2. Synthetic Dataset Construction and Fine-Tuning

After completing the optimization procedure, we assemble a synthetic dataset consisting of the final adversarial questions paired with expert-provided answers:

$$\mathcal{D}_{\text{synthetic}} = \left\{ \left(Q_i^{(T)}, f_{\text{strong}}(C, Q_i^{(T)}) \right) \right\}_{i=1}^{N}.$$

Fine-tuning the weak model f_{weak} on the expert-curated $\mathcal{D}_{\text{synthetic}}$ explicitly targets and remedies the shortcomings uncovered by adversarial optimization. Consequently, this focused training substantially enhances the model's robustness and accuracy on domain-specific comprehension tasks.

Algorithm 1 Iterative Adversarial Question Generation

Input: Domain context C; initial question set $\{Q_i^{(0)}\}_{i=1}^N$; strong model f_{strong}; weak model f_{weak}; feedback model f_{fb}; guidance model f_{guide}; revision model f_{rev}; number of iterations T
for $i = 1$ **to** N **do**
  for $t = 0$ **to** $T - 1$ **do**
    $A_{i,\text{strong}}^{(t)} \leftarrow f_{\text{strong}}(C, Q_i^{(t)})$
    $A_{i,\text{weak}}^{(t)} \leftarrow f_{\text{weak}}(C, Q_i^{(t)})$
    $\mathcal{L}(Q_i^{(t)}) \leftarrow f_{\text{fb}}(A_{i,\text{strong}}^{(t)}, A_{i,\text{weak}}^{(t)})$
    $\nabla_Q \mathcal{L}(Q_i^{(t)}) \leftarrow f_{\text{guide}}(Q_i^{(t)}, \mathcal{L}(Q_i^{(t)}))$
    $Q_i^{(t+1)} \leftarrow f_{\text{rev}}(Q_i^{(t)}, \nabla_Q \mathcal{L}(Q_i^{(t)}))$
  end for
end for
Return: $\{Q_i^{(T)}\}_{i=1}^N$ *(Final set of optimized questions)*

4. Experiments

We assess our method on a targeted subset of the **LegalBench** benchmark (Guha et al., 2023), which originally comprises a wide range of tasks intended to evaluate the legal reasoning abilities of large language models. To emphasize domain-specific abilities, we focus on the three most frequently referenced contracts from the CUAD dataset (Hendrycks et al., 2021) within the LegalBench suite. These are: *Cardlytics Maintenance Agreement*, *Buffalo Wild Wings Franchise Agreement*, and *PF Hospitality Franchise Agreement*. Across these contracts, there are a total of 491 benchmark questions spanning 36 distinct tasks.

Table 1. Accuracy (%) across three contract-specific subsets and average for LLaMA3-8b.

Method	Tokens	Cardl	Buffa	Pfhos	Avg
Baselines					
No Extra Data	-	67.3	69.1	72.1	69.5
Ours					
Ours	96k	**82.7**	**79.6**	**85.7**	**82.7**
Competitor					
Paraphrase $\times 6$	149k	68.5	70.4	77.0	71.9
Model-indep. QA	147k	75.0	74.1	78.3	75.8
Entigraph	6.7M	80.4	76.5	82.0	79.6
Knowledge_Instr	159k	78.6	70.4	75.8	75.0

We compare our proposed fine-tuning approach against a base pretrained LLM , as well as several increasingly sophisticated domain adaptation strategies. These include (1) paraphrase-based fine-tuning; (2) fine-tuning on uninformed questions, a naive synthetic dataset generated by prompting a strong LLM to write challenging questions for a given contract without any iterative refinement or adversarial objective; (3) EntiGraph-based augmentation (Yang et al., 2024b); and (4) Knowledge-Instruct (Ovadia et al., 2025). All models are evaluated using a few-shot setting with the standard LegalBench prompt template. To isolate the effects of fine-tuning strategies, we do not employ any retrieval augmentation or external tools.

Our adversarial QA generation approach consistently outperforms both paraphrase-based fine-tuning and model-independent questions fine-tuning baselines. It achieves a **18.99%** improvement over the base model and a **3.89%** gain relative to the EntiGraph competitor, while using $\approx 70\times$ fewer training tokens. Whereas EntiGraph primarily augments entity-level relational knowledge, our feedback-driven question optimization specifically targets interpretive shortcomings, leading to richer semantic coverage and greater task accuracy.

To assess the impact of iterative adversarial refinement, we perform a sensitivity analysis by varying the number of optimization steps during dataset construction. For each contract, we generate adversarial questions with different refinement iterations, fine-tune the target model, and evaluate on the corresponding LegalBench tasks.

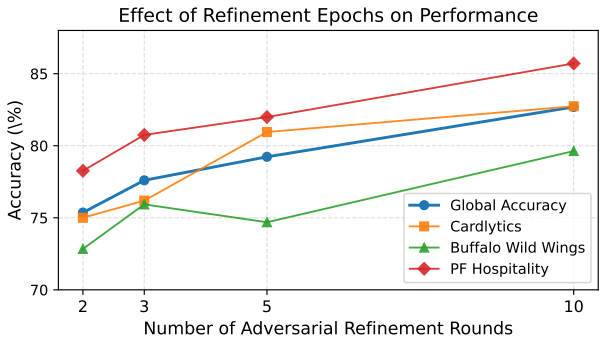

Figure 1. Accuracy on LegalBench (top three contracts) by epoch; performance improves with more refinement.

As shown in Figure 1, additional refinement iterations result in progressively higher accuracy, highlighting the effectiveness of iterative adversarial feedback in improving domain-specific model performance.

5. Conclusion

We introduce an adversarial question generation framework to enhance domain-specific LLMs by systematically identifying and addressing their interpretive weaknesses. By adapting the TextGrad differentiable prompting method, our approach iteratively produces targeted questions and synthetic datasets aligned with model-specific shortcomings. Empirical results demonstrate that this strategy substantially improves the reasoning performance of smaller LLMs, even in limited data settings, highlighting the effectiveness of adversarial refinement for efficient domain adaptation.

References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *CoRR*, abs/1810.08750, 2018. URL http://arxiv.org/abs/1810.08750.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/JCSS.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.740. URL https://doi.org/10.18653/v1/2020.acl-main.740.

Hendrycks, D., Burns, C., Chen, A., and Ball, S. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 328–339. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1031. URL https://aclanthology.org/P18-1031/.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. Tinybert: Distilling BERT for natural language understanding. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 4163–4174. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.372. URL https://doi.org/10.18653/v1/2020.findings-emnlp.372.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020. doi: 10.1093/BIOINFORMATICS/BTZ682. URL https://doi.org/10.1093/bioinformatics/btz682.

Lin, Y., Wang, C., Chang, C., and Sun, H. An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network. *Multim. Tools Appl.*, 80(3):4037–4051, 2021. doi: 10.1007/S11042-020-09276-9. URL https://doi.org/10.1007/s11042-020-09276-9.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Confere
html.

Maini, P., Seto, S., Bai, R. H., Grangier, D., Zhang, Y., and Jaitly, N. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 14044–14072. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.757. URL https://doi.org/10.18653/v1/2024.acl-long.757.

Ovadia, O., Brief, M., Mishaeli, M., and Elisha, O. Fine-tuning or retrieval? comparing knowledge injection in

llms. In Al-Onaizan, Y., Bansal, M., and Chen, Y. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 237–250. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.15.

Ovadia, O., Brief, M., Lemberg, R., and Sheetrit, E. Knowledge-instruct: Effective continual pre-training from limited data using instructions. *arXiv preprint arXiv:2504.05571*, 2025.

Ruder, S. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017. URL http://arxiv.org/abs/1706.05098.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Xu, Y., Sun, F., and Zhang, X. Literature survey of active learning in multimedia annotation and retrieval. In Lu, K., Mei, T., and Wu, X. (eds.), *International Conference on Internet Multimedia Computing and Service, ICIMCS '13, Huangshan, China - August 17 - 19, 2013*, pp. 237–242. ACM, 2013. doi: 10.1145/2499788.2499794. URL https://doi.org/10.1145/2499788.2499794.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=Bb4VGOWELI.

Yang, Z., Band, N., Li, S., Candès, E. J., and Hashimoto, T. Synthetic continued pretraining. *CoRR*, abs/2409.07431, 2024b. doi: 10.48550/ARXIV.2409.07431. URL https://doi.org/10.48550/arXiv.2409.07431.

Yüksekgönül, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., and Zou, J. Textgrad: Automatic "differentiation" via text. *CoRR*, abs/2406.07496, 2024. doi: 10.48550/ARXIV.2406.07496. URL https://doi.org/10.48550/arXiv.2406.07496.

A. Appendix

A.1. Detailed Code Explanation

This appendix provides an in-depth explanation of the implementation behind the adversarial question generation pipeline described in the main text. We outline its methodology, specific variable roles, code examples, and the adversarial optimization process.

A.2. Overall Methodology

The central goal of our approach is to iteratively generate adversarial questions that reveal and target the interpretive weaknesses of a smaller language model (termed the *weak model*), using discrepancies with a larger, more robust expert model (the *strong model*). The process relies on structured feedback from a feedback model, which evaluates the divergence between the models' responses in context.

Algorithm explanation: Algorithm explanation:

- *Inputs:* A set of N independently initialized questions (each starting from a placeholder, as described by the `question` variable in the next subsection), the fixed document context C, models f_{strong}, f_{weak}, and f_{fb}, as well as the total number of optimization iterations T.
- *Response Generation:* For each question Q_i at each iteration t, both models are queried with input $(C, Q_i^{(t)})$, yielding responses $A_{i,\text{strong}}^{(t)}$ and $A_{i,\text{weak}}^{(t)}$.
- *Discrepancy Evaluation:* The feedback model f_{fb} compares these responses, computing both a numeric disagreement score and a natural language explanation, denoted $\mathcal{L}(Q_i^{(t)})$.
- *Gradient Computation:* The guidance model f_{guide} receives the explanation from the feedback model and generates a natural language edit instruction, indicating how to revise $Q_i^{(t)}$ so as to further increase the answer discrepancy.
- *Question Update:* The revision model f_{rev} uses the current question and the edit instruction to generate an updated question, thereby steering the question to maximize divergence between the models' outputs.

After T rounds, the final set of optimized questions $\{Q_i^{(T)}\}_{i=1}^N$ together with their strong model answers are saved for constructing the synthetic dataset.

Implementation Note:
In practice, the variable `question` is initialized as a learnable object ('requires_grad=True') and is updated via feedback-driven optimization. All model weights and the context document remain fixed; only the content of the question is iteratively revised.

A.3. Implementation Details

Our implementation uses the `textgrad` framework, leveraging a differentiable prompting paradigm and modern LLM APIs. Below, each major component and step is described.

A.3.1. Model Configuration

Three key model interfaces are initialized via `LiteLLMEngine`:

- **Strong Model**: Provides authoritative answers strictly based on the contract. In practice, we use gpt-4o-mini.
- **Weak Model**: The target of improvement, typically smaller or less domain-specialized. In practice, we use llama3.1:8b.
- **Feedback Model**: Evaluates discrepancies between strong and weak model answers to guide question optimization.

A.3.2. Central Role of the Optimized `question` Variable

A distinctive and central aspect of our method is that the `question` variable itself is the object of optimization. Instead of tuning model parameters, our pipeline holds both the contract text and all model weights fixed—and iteratively refines the natural language content of the `question` prompt.

In code, the variable is initialized as:

```
question = tg.Variable(
"Q: ???",  # minimal placeholder
requires_grad=True,
role_description="A legal
question to be optimized for
maximal response divergence"
)
```

Conceptually, at each iteration, the current value of `question` is updated to maximize measured discrepancy between the weak and strong models. This iterative process adaptively generates natural language questions that are adversarial: they are optimized to be especially likely to expose errors, misunderstanding, or blind spots in the weak model, given a fixed contract context.

A.3.3. Prompt Templates

Carefully crafted prompts ensure the models fulfill their roles.

Strong and Weak Models Prompt:

```
You are an expert in
interpreting contracts.
The user will provide a contract
and a question about it.
Provide the most accurate,
thorough answer based on the
contract's text.
Stay strictly within the
contract's details and do not
invent external laws.
```

Feedback Model Prompt:

```
You are an expert in legal
contract analysis.
Given two responses (correct vs.
potentially incorrect), identify
contradictions, omissions, or
errors in the second response.
Provide a numeric incorrectness
score (0.0 to 1.0) with detailed
explanation.
```

Example formatted prompt for discrepancy evaluation:

```
Compare the following responses:
<CONTRACT>: {contract}
<QUESTION>: {question}
<CORRECT_RESPONSE>:
{response_strong}
<POSSIBLY_INCORRECT_RESPONSE>:
{response_weak}
Rate the incorrectness of the
second response (0 to 1) and
explain errors, contradictions,
or missing details.
```

A.3.4. Optimization Algorithm

Optimization employs the TextGrad framework's differentiable prompting technique. Each iteration refines `question` based on textual feedback, with code logic as follows:

```
optimizer.zero_grad()
divergence_loss =
loss_fn(contract_text, question,
response_strong, response_weak)
divergence_loss.backward()
optimizer.step()
```

This iterative update is *conceptually* analogous to a gradient ascent step, since each feedback-driven edit aims to increase (rather than decrease) the disagreement proxy—but the updates are performed purely in the space of natural language.

A.4. Synthetic Dataset Construction

After optimization, synthetic adversarial questions and their authoritative answers (from the strong model) are saved to a dataset. This dataset is constructed specifically to target and address the interpretative weaknesses surfaced in the weak model.

A.5. Illustrative Example: Iterative Adversarial Refinement of Legal Questions

To concretely illustrate our feedback-driven methodology, we present an example drawn from adversarial QA generation over a complex legal contract—the "Software License, Customization and Maintenance Agreement" between Bank of America and Cardlytics, Inc. At each iteration t, our algorithm interacts with the document context C, refines the current question $Q^{(t)}$, and leverages feedback from model disagreements to systematically probe where the target model falls short.

Initial Setup

- **Document Context (C):** Raw, multi-page legal contract (e.g., 30+ pages covering IP, confidentiality, audit, etc.).
- **Current Question $Q^{(t)}$:** The natural-language prompt being optimized.
- **Strong (Expert) Model:** Large LLM (e.g., GPT-4o-mini in our experiments) with robust reasoning over C.
- **Weak (Target) Model:** Smaller LLM (e.g., Llama-3 8B in our experiments); less accurate on nuanced legal reasoning.
- **Feedback Model:** Compares strong/weak responses, assigns a numerical error score, and provides actionable natural-language critique highlighting omissions, misconceptions, or interpretive errors.

Iteration 0: Initialization $\mathbf{Q^{(0)}}$: `Q: ???`

Strong Model's Response: *"It appears that you did not provide a specific question regarding the contract. Please clarify your question, and I will do my best to provide an accurate and thorough answer based on the contract's text."*

Weak Model's Response: *"... it appears that you are asking about a specific aspect... If I were to guess, one possible question could be: 'What are the requirements for background checks...'..."*

Feedback Model Output:

- **Score:** 0.5 (partially incorrect/misleading)
- **Critique:** Weak model guesses rather than prompting for clarification; fails to require user input.
- **TextGrad Guidance:** "Replace vague placeholder with a meaningful, contextually grounded question targeting a specific clause (e.g., confidentiality, indemnity). Use hypotheticals or contrast with industry standards."

Revision $\rightarrow$ $\mathbf{Q}^{(1)}$: `What are the potential risks and benefits associated with the confidentiality provisions in this contract, and how might these affect both parties in terms of liability and operational flexibility?`

Iteration 1: Sharpening Scope and Depth Strong Model: Multi-paragraph answer, citing specific sections, discussing operational and compliance implications.

Weak Model: Discusses generic risks/benefits but omits contract-specific definitions and scenario detail.

Feedback Model Output:

- **Score:** 0.6 (incomplete)
- **Critique:** Omits key details (e.g., definition of "confidential info", obligations on breach, liability nuances).
- **TextGrad Guidance:** "Expand comparative scope—request scenario analysis, stakeholder view, and cross-industry contrast (e.g., healthcare, GDPR)."

Revision $\rightarrow$ $\mathbf{Q}^{(2)}$: `Considering the entire Agreement, what are the potential risks and benefits associated with the confidentiality provisions, and how do these provisions compare to those in other financial services contracts? Can you provide examples or scenario analyses for their impact on liability and operational flexibility for both parties and stakeholders involved?`

Iteration 2: Multi-Perspective and Hypothetical Analysis Strong Model: Expands with cross-industry contrasts (e.g., with HIPAA), stakeholder-specific discussion, hypothetical adverse event scenarios.

Weak Model: Improves, but still superficial on scenario depth and long-term effects.

Feedback Model Output:

- **Score:** 0.6
- **Critique:** Insufficient detail in long-term consequences, stakeholder-specific risk perception.
- **TextGrad Guidance:** "Ask for best- vs. worst-case hypotheticals, speculative future impact, and divergent interpretations by different actors."

Revision $\rightarrow$ $\mathbf{Q}^{(3)}$: `How do the confidentiality provisions, liability clauses, and compliance requirements in this agreement influence operational flexibility and stakeholder perspectives? How do they compare to industry standards across healthcare, technology, and privacy? Provide contrasting hypothetical scenarios and discuss possible long-term consequences of breaches or compliance failures (including differing stakeholder interpretations).`

Final feedback at T iteration This adversarial, iterative loop yields a maximally challenging question, such as:

`In what ways do the confidentiality provisions, liability clauses, and compliance requirements influence operational flexibility and stakeholder perspectives? How do these elements compare with industry standards across sectors such as healthcare and technology? Can you provide contrasting hypothetical scenarios showing potential risks and benefits, with a future-facing discussion on how evolving laws and technologies might force adaptation? What differing interpretations might various stakeholders bring to these elements?`