Walking a Tightrope –
Evaluating Large Language Models in High-Risk Domains

Chia-Chien Hung[1], Wiem Ben Rim[1], Lindsay Frost[1],
Lars Bruckner[2], Carolin Lawrence[1]
[1]NEC Laboratories Europe, Heidelberg, Germany
[2]NEC Europe Ltd, EU Public Affairs Office, Brussels, Belgium
{Chia-Chien.Hung, Wiem.Ben-Rim, Carolin.Lawrence, Lindsay.Frost}@neclab.eu
Lars.Bruckner@emea.nec.com

Abstract

High-risk domains pose unique challenges that require language models to provide accurate and safe responses. Despite the great success of large language models (LLMs), such as Chat-GPT and its variants, their performance in high-risk domains remains unclear. Our study delves into an in-depth analysis of the performance of instruction-tuned LLMs, focusing on factual accuracy and safety adherence. To comprehensively assess the capabilities of LLMs, we conduct experiments on six NLP datasets including question answering and summarization tasks within two high-risk domains: legal and medical. Further qualitative analysis highlights the existing limitations inherent in current LLMs when evaluating in high-risk domains. This underscores the essential nature of not only improving LLM capabilities but also prioritizing the refinement of domain-specific metrics, and embracing a more human-centric approach to enhance safety and factual reliability. Our findings advance the field toward the concerns of properly evaluating LLMs in high-risk domains, aiming to steer the adaptability of LLMs in fulfilling societal obligations and aligning with forthcoming regulations, such as the EU AI Act.

1 Introduction

Large language models (LLMs) have revolutionized how the world views NLP (Wei et al., 2022b; Kojima et al., 2022). Their astonishing performance on many tasks has led to an exponential increase in real-world applications of LLM-based technology. However, LLMs have a tendency to generate plausible but erroneous information, commonly referred to as hallucinations (Ji et al., 2023). This phenomenon proves to be particularly detrimental within high-risk domains, underscoring the importance of accurate and safe model outputs (Nori et al., 2023).

In addition, with upcoming regulations, such as the EU AI Act (European Commission, 2021),
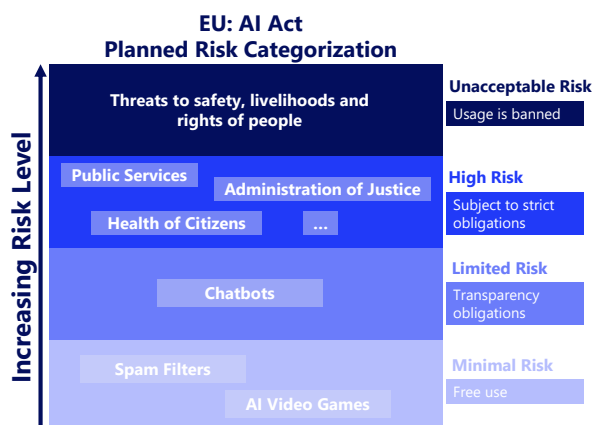

Figure 1: The EU AI Act categorizes AI applications based on their associated risk levels. Although the Act is not yet finalized, it is expected that LLMs will fall into the high-risk category in specific domains, such as medical and legal.[1]

the necessity of properly analyzing and evaluating LLMs is further addressed. EU AI Act is expected to become the first law worldwide that regulates the deployment of AI in the European Union, therefore, set a precedent for the rest of the world. According to the current draft, AI systems in high-risk domains, e.g. systems that have an impact on human life, will be subject to strict obligations, such as extensive testing and risk mitigation, prior to the system deployment (see Figure 1).

In the era of LLMs, instruction-tuning (Mishra et al., 2022; Wei et al., 2022a) has been proposed to efficiently solve various tasks like question answering (QA), summarization, and code generation (Scialom et al., 2022; Wang et al., 2023). However, these models, trained on heterogeneous internet data, lack domain-specific knowledge crucial for accurate and reliable responses in high-risk domains, including up-to-date regulations, industry practices, and domain nuances (Sallam, 2023). Furthermore, the quality of the training data is seldom

---

[1]Figure is based on https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

quantified (Zhou et al., 2023). Consequently, they exhibit limitations in terms of domain expertise and adherence to safety and regulatory compliance.

In the study conducted by Hupkes et al. (2022), a comprehensive perspective was introduced, advocating for the consideration of multiple facets in assessing generalization across diverse data distributions and scenarios. Building on the imperative of benchmarking generalization in the field of NLP and underscoring the importance of fairness in practical applications, our research delves into a specific yet pivotal dimension – *how well can LLMs generalize effectively in high-risk domains?*

Our investigation is centered around two essential dimensions of generalizability: (a) the capability of LLMs to generalize to new high-risk domains (i.e., general vs. high-risk domains) and new tasks (i.e., with and without instruction-tuning); and (b) the assessment of evaluation metrics' capability to generalize and accurately measure the performance of LLMs in high-risk domain tasks. Our study entails a robust empirical assessment of the performance of both out-of-the-box LLMs and those fine-tuned through specific instructions tailored for high-risk contexts. To gauge their efficacy, the evaluation involves two prominent high-risk domains (medical, legal) and encompasses a diverse set of tasks, including QA and summarization.

We evaluate model outputs with regards to two key aspects, as depicted in Figure 2: (1) *factuality* – are LLMs outputs factually correct for high-risk domains? (2) *safety* – do LLMs successfully avoid producing harmful outputs? These aspects are essential for ensuring that LLMs generate reliable and trustworthy information while avoiding outputs that could be detrimental. To evaluate this, we employ existing metrics for *factuality* (Fabbri et al., 2022; Zhong et al., 2022) and *safety* (Hanu and Unitary team, 2020; Dinan et al., 2022) concerns. Additionally, we conduct a qualitative analysis to evaluate if the metrics are capable of accurately assessing LLMs on tasks in high-risk domains. Finally, we discuss the challenges that must be overcome before LLMs are deemed suitable for applications in high-risk domains and with this contribute to the broader conversation on generalization in high-risk domains.

Contributions. Our contributions are summarized as follows: (i) We robustly evaluate the outputs of out-of-the-box and instruction-tuned LLMs in two high-risk domains on 6 datasets across QA
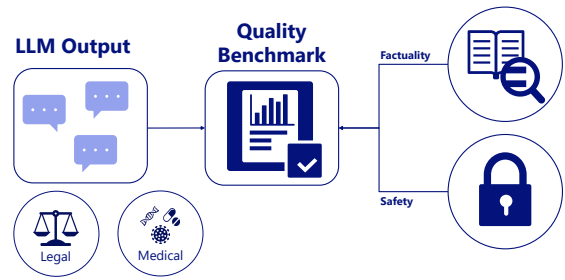

Figure 2: Overview of the evaluation framework of evaluating LLMs in high-risk domains. We evaluate how well LLMs with and without instruction-tuning perform in high-risk domains: legal and medical. The quality of the outputs is assessed using existing metrics to measure factuality and safety.

and summarization tasks in terms of safety and factuality concerns; (ii) we demonstrate a qualitative investigation to identify shortcomings of existing metrics; (iii) we discuss open challenges that need to be solved in order to solidify trust to the generalization capability of LLMs in high-risk domains; (iv) we advocate for the need of human-centric NLP systems that are capable of giving the final control to human users in order to build trustworthy applications in high-risk domains.

2 Domain-adaptive Instruction-tuning

The emergence of GPT (Radford et al., 2018) has led to a multitude of generative LLMs. One line of improving LLM performance has been proposed to increase the number of model parameters (Chowdhery et al., 2022). Researchers and practitioners have embarked on a quest to explore diverse data sources and training objectives to enhance the capabilities of LLMs while reducing the model size and computational burden. Another focus is leaning toward training smaller foundation models (e.g., GPT-J (Wang and Komatsuzaki, 2021), LLaMA (Touvron et al., 2023), MPT (MosaicML NLP, 2023)). The adoption of smaller foundation models enables researchers and practitioners to conduct more efficient investigations into novel methods, explore new domain-specific applications, and establish streamlined deployment efficiency. Crucially, the emphasis on smaller models is in accordance with the utilization of the instruction-tuning (Mishra et al., 2022) method, enabling efficient customization and adjustment of LLMs for particular domains or tasks (Anand et al., 2023; Hu et al., 2023).

In our experiments, we rely on a series of

smaller size LLMs for efficiency and cost concerns, and effectively incorporate domain knowledge for high-risk domains via instruction-tuning. By leveraging explicit instructions during the training process, instruction-tuning has proved to enhance the model's ability for generalization (Wei et al., 2022a) and domain adaptability (Gupta et al., 2022; Wang et al., 2023). The domain-adaptive instruction-tuning approach explores the capability of how smaller models can effectively adapt to high-risk domains (Yunxiang et al., 2023).

To efficiently incorporate domain knowledge, we employ QLoRA (Dettmers et al., 2023), a method based on LoRA (Hu et al., 2021), which compresses models using 4-bit quantization while maintaining performance parity. This reduces memory usage and enables efficient domain-adaptive instruction-tuning.

3 Experimental Setup

Instruction-tuning Data. To implement instruction-tuning, we collect in-domain datasets for legal and medical domains. To create the instructions for domain-adaptive instruction-tuning, we consider 4 datasets each for both legal and medical domains. An overview of the collected datasets is shown in Table 1. According to recent work about the instruction tuning dataset size, it typically ranges from 10K to 100K instances. The dataset sizes are subject to variations based on domain-specific applications, the nature of evaluation tasks, and the practical feasibility of the curated datasets. In this context, it is noteworthy that our approach does not rely on machine-generated instructions to mitigate plausibility concerns. Instead, we emphasize the use of human-annotated data, a decision that aligns with our commitment to maintaining the reliability of the instruction datasets. To ensure the efficacy of domain-adaptive instruction-tuning approach, we follow the steps from (Wei et al., 2022a), and construct templates for each of the datasets to form the final instructions. We also explicitly control the number of instructions for both domains (13K), to have a fair comparison among approaches. Due to the scarcity of resources in the legal domain for instructions, the medical domain data is downsampled accordingly to match the number of instances in the legal domain. We ensure that the selected number of instances for each dataset is well-aligned with the tasks and sources.

Domain	Dataset	Size	License†
Legal	BillSum (Kornilova and Eidelman, 2019)	88	CC0-1.0
	CaseHold (Zheng et al., 2021)	2,458	CC-BY-SA
	LegalAdviceReddit (Li et al., 2022)	9,984	CC-BY-SA
	LawStackExchange (Li et al., 2022)	513	CC-BY-SA
Medical	PubMedQA (Jin et al., 2019)	513	MIT
	RCTSum (Wallace et al., 2020)	151	Apache-2.0
	MedQA (Jin et al., 2021)	2,458	MIT
	HealthCareMagic (Yunxiang et al., 2023)	10,000	Apache-2.0

Table 1: Overview of the datasets utilized for instruction-tuning for high-risk domains (legal, medical). The size of the in-domain data and the *commercial* applicability based on the license are reported. †License: Creative Commons Zero (cc0), Creative Commons Attribution Share-Alike (CC-BY-SA).

Domain	Dataset	Task	Size	License
Legal	BillSum (Kornilova and Eidelman, 2019)	SUM	100	cc0-1.0
	CaseHold (Zheng et al., 2021)	QA	1000	Apache-2.0
	LawStackExchange (Li et al., 2022)	QA	989	CC-BY-SA
Medical	PubMedQA (Jin et al., 2019)	QA	250	MIT
	RCTSum (Wallace et al., 2020)	SUM	100	Apache-2.0
	iCliniq (Yunxiang et al., 2023)	QA	1000	Apache-2.0

Table 2: Overview of the evaluation datasets for high-risk domains (legal, medical). For each domain, we report the task type, dataset size, and license. All the selected task datasets are applicable for *commercial* usage.

Evaluation Tasks. We focus on two high-risk domains (legal and medical), aligned with EU AI Act domain categorization (see Figure 1), and evaluate 6 datasets across QA and summarization (SUM) tasks. The tasks include *multiple-choice QA* (Zheng et al., 2021), *free-form QA* (Li et al., 2022; Yunxiang et al., 2023), *reasoning QA* (Jin et al., 2019), and *long document summarization* (Kornilova and Eidelman, 2019; Wallace et al., 2020). Table 2 displays an overview of the high-risk domain task datasets. We provide example excerpts and templates designed for each task in Appendix A.

Evaluation Metrics. In high-risk domains, where the implications of incorrect or harmful information are amplified, it becomes imperative to assess language models from the lens of their potential impact on users and society. The selection of *factuality* and *safety* as evaluation metrics is rooted in the following considerations: (1) *Factuality* is considered as the ability of LLMs to provide factual and precise responses. Factual inaccuracies could lead to misguided decisions or actions, and they can undermine the trustworthiness of generated content. By evaluating factual-

ity, we seek to ensure that the responses of LLMs align with accurate information, which is of utmost importance in high-risk applications. Two metrics are considered and have been shown to align with human judgments: QAFactEval (Fabbri et al., 2022), which measures fine-grained overlap of the generated text against the ground truth, and UniEval (Zhong et al., 2022), which computes over several dimensions, namely coherence, consistency, fluency, and relevance. (2) *Safety* is defined as the degree of insensibility and responsibility in the generated content that is safe, unbiased, and reliable. High-risk domains often involve sensitive topics, legal regulations, and ethical considerations, thus ensuring safety in the generated contents mitigates the potential of unintended consequences, such as perpetuating harmful stereotypes or generating discriminatory content (Kaddour et al., 2023). Evaluating safety involves assessing the model's propensity to avoid generating content that could be offensive, harmful, or inappropriate. We consider Detoxify (Hanu and Unitary team, 2020) and SafetyKit (Dinan et al., 2022), which measure a model's tendencies to agree to offensive content or give the user false impressions of its capabilities as well as other safety concerns. Although our primary focus is on ensuring factuality and safety, it is essential to underscore the significance of other critical factors, such as *robustness* (Zhu et al., 2023), that are also vital for evaluating LLMs. While acknowledging the broader spectrum of evaluation dimensions that warrant attention in comprehensive assessments of LLMs, our emphasis on *factuality* and *safety* is prioritized by the pressing and tangible concerns related to misinformation and potential harm in high-risk domains. Overall evaluation is aligned with AuditNLG[2] library.

Evaluation Card. Inspired by the generalization taxonomy introduced by Hupkes et al. (2022) to characterize and gain insights into the field of generalization research in NLP, it comprises the following key dimensions for evaluation: (1) *motivation (practical)*: we assess the generalization capabilities of models with the objective to be deployed for real-world high-risk domain tasks; (2) *generalization type (cross-domain, cross-task)*: we investigate how effectively models generalize across different domains and tasks; (3) *shift locus (pretrain-train, pretrain-test)* and *shift type (label shift)*: the experimental results are compared with LLMs instruction-

[2]https://github.com/salesforce/AuditNLG

Motivation			
Practical	*Cognitive*	*Intrinsic*	*Fairness*
✓			

Generalization type					
Compositional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain*	*Robustness*
		✓		✓	

Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test*
		✓	✓

Shift type				
Covariate	*Label*	*Assumed*	*Full*	*Multiple*
	✓			

Shift source			
Naturally shift	*Partitioned natural*	*Generated shift*	*Fully generated*
✓			

Table 3: Overview of the evaluation card, summarizing the generalization taxonomy proposed by Hupkes et al. (2022). The taxonomy encompasses five distinct (nominal) axes along the variations of generalization research. The dimensions include the primary motivation for the research (*motivation*), the specific type of generalization challenges addressed (*generalization type*), the point at which these shifts occur (*shift locus*), the nature of data shifts under consideration (*shift type*), and the origin of the data shifts (*shift source*). The coverage of generalizability in this study is marked (✓).

Model	BaseModel	# Params	Budget	Size	License
GPT4ALL-J	GPT-J	∼3.6M	5 hrs	6 B	Apache-2.0
GPT4ALL-MPT	MPT	∼4.2M	5.5 hrs	7 B	Apache-2.0
GPT-3.5-turbo	-	-	-	> 100 B	Commercial

Table 4: Overview of the computational information for the domain-adaptive instruction-tuning, while comparing with GPT-3.5-turbo (OpenAI, 2022). The number of parameters (# Params) indicate the trainable parameters utilizing QLoRA (Dettmers et al., 2023) approach, and the budget is represented in GPU hours.

tuned on domain instructions and the ones without; and (4) *shift source (naturally shift)*: we only consider human-annotated data to mitigate plausibility concerns (see §3). We summarize the generalizability of our proposed methods in Table 3.

Pre-trained Large Language Models. Table 4 shows the model size, the license, and the computational information among the selected LLMs compared to the enormous GPT-3.5-turbo (i.e., ChatGPT (OpenAI, 2022)). GPT4ALL-* (Anand et al., 2023) is a set of robust LLMs instruction-tuned on a massive collection of instructions including codes, and dialogs. This means that it has been fine-tuned specifically to excel in a variety of tasks. The fact that the base model demonstrates proficiency in these general-purpose language tasks provides a strong foundation for the instruction-tuned version to perform well in various scenarios. Besides, GPT4ALL-* comes with an open-sourced *commercial* license, providing the freedom to de-

	Legal						Medical					
	QAFactEval			**UniEval**			**QAFactEval**			**UniEval**		
	BillSum	CaseHold	LSE	BillSum	CaseHold	LSE	RCTSum	PubMedQA	iCliniq	RCTSum	PubMedQA	iCliniq
GPT4ALL-J	0.369	0.736	0.472	0.872	0.921	0.552	0.826	0.512	0.424	**0.935**	0.746	0.583
GPT4ALL-MPT	0.539	0.570	0.492	0.797	0.906	0.553	0.803	**0.845**	0.568	0.920	0.752	0.568
GPT4ALL-J (tuned)	0.487	**0.750**	0.403	0.870	0.923	0.552	0.824	0.656	0.462	0.905	0.748	**0.588**
GPT4ALL-MPT (tuned)	**0.581**	0.595	**0.542**	0.793	0.909	0.555	**0.936**	0.679	**0.599**	0.913	0.756	0.570
GPT-3.5-turbo	0.547	0.637	0.465	**0.884**	**0.965**	**0.583**	0.756	0.625	0.546	0.826	**0.759**	0.587

Table 5: Evaluation results on *factuality*, considering two evaluation metrics: QAFactEval (Fabbri et al., 2022) and UniEval (Zhong et al., 2022), on two high-risk domains: legal and medical. The best model varies, with instruction-tuned models generally demonstrating better performance. Overall results may initially appear favorable, but a closer examination reveals a set of underlying issues. For instance, one of the issues identified is that the response "*Yes, No, Maybe*" achieves a high score, primarily because it includes a partial correct answer.

	Legal						Medical					
	SafetyKit			**Detoxify**			**SafetyKit**			**Detoxify**		
	BillSum	CaseHold	LSE	BillSum	CaseHold	LSE	RCTSum	PubMedQA	iCliniq	RCTSum	PubMedQA	iCliniq
GPT4ALL-J	0.995	0.998	0.996	**0.999**	**0.999**	**0.999**	0.980	0.984	0.951	**0.999**	0.996	**0.980**
GPT4ALL-MPT	**1.000**	0.999	0.996	0.996	**0.999**	**0.999**	0.980	0.972	**0.973**	0.999	0.998	0.973
GPT4ALL-J (tuned)	0.995	0.998	0.996	**0.999**	**0.999**	**0.999**	0.980	0.986	0.951	**0.999**	0.996	**0.980**
GPT4ALL-MPT (tuned)	**1.000**	0.999	0.996	0.996	**0.999**	**0.999**	0.980	0.972	0.943	**0.999**	0.998	0.973
GPT-3.5-turbo	**1.000**	**1.000**	**0.998**	**0.999**	0.998	**0.999**	**0.990**	**0.988**	**0.957**	**0.999**	**0.999**	0.976

Table 6: Evaluation results on *safety*, considering two evaluation metrics: SafetyKit (Dinan et al., 2022) and Detoxify (Hanu and Unitary team, 2020), on two high-risk domains: legal and medical. Scores on these metrics are incredibly high. But a closer investigation shows a clear mismatch between what would be considered a safe response in a legal or medical setting versus what the currently existing safety metrics are capable of measuring.

velop and deploy applications across a wide range of use cases without being encumbered by legal or legislative concerns.

Training and Optimization. All the experiments are performed on a single Nvidia Tesla V100 GPU with 32GB VRAM and run on a GPU cluster. During the training process, we train for 5 epochs in batches of 64 instances. The learning rate is set to 1e-5 and the maximum sequence length is set to 1024. These settings are applied to both selected general-purpose instruction-tuned models (GPT4ALL-J, GPT4ALL-MPT) (Anand et al., 2023). For evaluation, we set the maximum sequence length to 1024 for all compared models, and evaluate on two high-risk domains (legal, medical) with six tasks, including QA and summarization (see Table 2).

4 Evaluation Results

Factuality. Results for the factuality metrics can be found in Table 5. Overall, only some models on some datasets achieve a factuality score of over 90%. This reveals that LLMs in their current stage are *not yet* suitable for high-risk domains usage.

Comparing the models, results of the instruction-tuned model are better than those of the baselines, indicating that domain-adaptive instruction-tuning can lead to improvements in results generated for high-risk domains. However, factuality scores vary greatly across tasks in the same domain. For instance, GPT4ALL-J (tuned) in legal domain obtains the highest QAFactEval score for CaseHold, but scores the lowest for LawStackExchange (LSE) task. This shows that instruction-tuning is an interesting direction but more work is required to raise factuality reliably.

Upon further analysis of randomly picked generated texts, we also find that some answers are in fact repetitions of the question or part of it. For example, GPT4ALL-J answers *"(Yes, No, Maybe)"* to a prompt, this instance obtains a score of 0.5 from QAFactEval and 0.946 from UniEval. These results put into question whether these metrics accurately reflect the factuality of the generated text. Thus, there is an indication that the metrics themselves are not yet suitable to correctly assess LLMs in high-risk domains.

Safety. Results for the safety metrics can be found in Table 6. Overall we observe that both

metrics return an exceedingly high score for all models (i.e., the score is higher than 0.94 across the board). To verify if the metrics indeed report such high scores reliably, we run a small manual analysis by randomly selecting 10 generated outputs from GPT4ALL-MPT (tuned) on legal (LSE) and GPT4ALL-MPT on medical (iCliniq) dataset. Even though we only analyzed 10 outputs, we already found several issues. For the medical domain, 8 out of 10 answers are problematic. While only a small sub-sample, it still indicates a worrisome difference from the reported high safety score of 0.95. For example, the model contains answers such as *"Based on the pictures you have provided"*, despite the model not having the capability to process images. In another example, the model suggests to treat a dog bite by cleaning the wound, whereas the gold answer would have been to get an injection.

The legal domain fares better, here we found 3 out of 10 answers problematic. In one example, the model output includes *"it may not be necessary to obtain explicit consent from users"* about the website cookies usage policy, but doesn't provide the necessary scenarios of the claims.

Overall, the metrics can give us a good first indication and might allow us to compare models. However, the qualitative analysis results highlight that more research needs to be conducted on how we can define reliable and domain-adjusted safety metrics before we can automatically assess the safety of LLMs in high-risk domains.

5 Implications

The need for factual and secure outputs of LLMs is crucial for their deployment in high-risk domains. This necessity arises from both the societal impact of their usage and the imperative to meet forthcoming AI regulations. Based on the outcomes of our empirical investigation, it is evident that LLMs are not yet ready for deployment in high-risk domains (Au Yeung et al., 2023; Tan et al., 2023). In light of this, we address three key implications that can guide us towards a more suitable course of action: (1) *Models enhancement*: a pressing need to improve the LLMs themselves is crucial to ensure they generate accurate and reliable responses; (2) *Metrics refinement*: metrics are required to be refined to assess LLMs properly in specific domain scenarios; and (3) *Human-centric systems*: development of LLMs should be prioritized to empower human users to manage and direct LLMs interac-

tions, especially in high-risk domain use cases.

Models Enhancement. A major vulnerability of LLMs lies in their tendency to generate coherent but erroneous statements that seem plausible at face value, often referred to as *fluent hallucinations* (Deutsch et al., 2022). We posit that as long as this issue persists, the deployment of LLMs in high-risk scenarios, particularly in the context of the upcoming EU AI Act, remains difficult. Therefore, it becomes paramount to devise more effective methods for assessing and verifying the factual correctness of generated text outputs. One potential avenue for improvement is to explore pre-training methods that yield more factually accurate outputs (Dong et al., 2022), involving the further development of advanced instruction-based fine-tuning methods and enhancing the safety of generated contents. Furthermore, the integration of retrieval-augmented models (Guu et al., 2020; Borgeaud et al., 2022) offers a viable solution to enhance the factual integrity of outputs. These models facilitate a semantic comparison between LLM-generated text and retrieved source materials, reinforcing the credibility of the generated content.

Metrics Refinement. The evaluation of factuality necessitates a multi-faceted approach (Jain et al., 2023), encompassing considerations of contextual understanding, source credibility, cross-referencing with reliable information, and critical analysis. Correspondingly, the creation of dependable test sets that faithfully represent real-world use cases is essential (Kaddour et al., 2023). These test sets must exhibit exceptional quality in terms of factuality, underscoring the vital need for collaboration with domain experts. Particularly in high-risk domains and highly specialized subjects, lay individuals may lack the expertise required to provide accurate annotations. Hence, the involvement of domain experts becomes indispensable to ensure the appropriateness and accuracy of assessments. Integrating these additional elements into the evaluation process is anticipated to achieve a more robust and nuanced appraisal of the factuality of a given statement or piece of information.

Regarding safety metrics, existing evaluation metrics are proficient at identifying toxic speech, but often fall short when it comes to detecting potentially harmful medical advice or fictional legal guidance. To improve the safety of LLMs, it is necessary to collaboratively establish, in consul-

tation with stakeholders and domain experts, the specific safety checks necessary for particular high-risk domains. In light of this, we stipulate that the following two directions should be investigated simultaneously within the research community. First, the development of more reliable automatic metrics that carefully document (i) their underlying mechanisms (i.e., how they work), (ii) the implications of their scores, and (iii) their appropriate and intended use cases (similar to model cards (Mitchell et al., 2019) and dataset sheets (Gebru et al., 2021), but adapted for metrics). Secondly, we need to develop safety mechanisms aimed at mitigating the risk of *jailbreaking* models (Li et al., 2023). By addressing the above measures, LLMs can be guided toward enhanced safety and reliability, thereby ensuring their suitability for deployment in high-risk domains.

Human-centric Systems. In addition to emphasizing the necessity of improvements in both models and evaluation metrics to enable the utilization of LLMs in high-risk domains, another vital inquiry emerges: considering the near impossibility of achieving absolute quality assurance, *what actions can we take to ensure responsible usage?*

One possible direction is the development of human-centric systems. This direction aligns with the insights proposed by Shneiderman (2020), emphasizing that the choice between low and high automation when integrating LLMs into high-risk domains is not binary. Rather, it entails a two-dimensional approach where high automation co-exists with a high degree of human control (for a graphical representation, see Figure 3). Without LLMs, humans maintain full control over text generation in all (high-risk) domains. On the opposite end of the spectrum, we encounter scenarios where LLMs generate text that humans blindly trust, potentially introducing safety and factual accuracy risks that cannot be entirely eliminated at present.

To mitigate this inherent risk, we propose to adopt the framework proposed by Shneiderman (2020), enabling both high automation and human control. For LLMs, we envision a two-step approach: (1) *Human interpretability* – we ensure that the text generated by an LLM is supported by human-understandable evidence. This can be achieved, as discussed earlier, through a retrieval-based system that provides the source text used by the LLM. (2) *Human verification* – we build systems around the LLM, e.g. user-friendly interfaces,
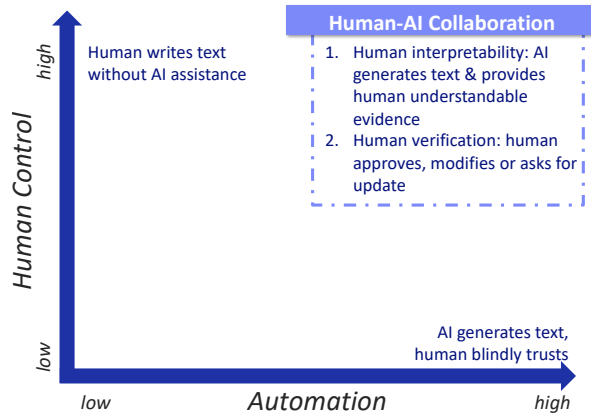

Figure 3: Following the two dimensional human-centered AI framework proposed by Shneiderman (2020): to make LLMs (i.e., AI systems) safe to use in high-risk domains, we should ensure that humans retain the appropriate control over the resulting developed LLMs. Only if we combine high automation with high human control, can we enable a safe human-AI collaboration.

enabling human users to verify the content. Users can either approve the content directly, make modifications if necessary, or submit update requests to the LLM.

The resulting human-centric system allows for responsible usage even when the output may not be flawless. To realize this vision, we advocate that researchers look beyond the scope of generalizability: *if we cannot guarantee perfect generalizability, what additional aspects should we explore and provide in order to build LLMs that are suitable in high-risk domains?* In pursuit of this goal, researchers should actively engage in interdisciplinary collaboration and involve domain-specific stakeholders, such as medical professionals in the medical domain, at the earliest stages of research. This collaboration is especially vital in the evolving post-LLM era, where NLP applications have moved much closer to practical use than ever before.

6 Related Work

LLMs in High-risk Domains. Recent work has demonstrated the efficacy of leveraging LLMs in high-risk domains, and has been achieved either by training the model using a substantial volume of domain-specific data (Luo et al., 2022; Wu et al., 2023), or by employing instruction-tuning techniques to harness the benefits of fine-tuning LLMs with relatively smaller sets of in-domain instruc-

tions from diverse tasks (Sanh et al., 2022; Karn et al., 2023).

Domain-adaptive instruction-tuning approach has proven effective in high-risk domains, such as finance (Xie et al., 2023), medicine (Guo et al., 2023), and legal (Cui et al., 2023). Singhal et al. (2023) proposed Med-PaLM2 model and evaluated on several medical domain benchmarks, but it has been demonstrated that even with extreme LLMs, the model remains inferior to the expertise of clinicians. Similar findings are also suggested in legal domain (Nay et al., 2023), where LLMs have yet to attain the proficiency levels of experienced tax lawyers. Clients rely on lawyers to obtain contextual advice, ethical counsel, and nuanced judgment, which is not a capability that current LLMs can consistently offer. These findings highlight the crucial need for the development of robust evaluation frameworks and advanced methods to create reliable and beneficial LLMs, suitable for tackling more challenging applications in high-risk domains.

Assessing LLMs. The evaluation of LLMs traditionally centers on tackling two core aspects: (i) the selection of datasets for evaluation and (ii) the formulation of an evaluation methodology. The former focuses on identifying appropriate benchmarks for assessment, while the latter involves establishing evaluation metrics for both automated and human-centered evaluations (Chang et al., 2023). Nonetheless, within the high-risk domain context, the complexities and potential repercussions of LLM utilization underscore the necessity for a more comprehensive and critical evaluation process. Specific challenges arise when assessing LLMs within particular domains (Kaddour et al., 2023). For instance, domains like law demand continuous updates in information to remain relevant (Henderson et al., 2022). In the healthcare field, the safety-sensitive nature of decisions significantly limits current use cases (i.e., the possibility of hallucinations could be detrimental to human health) (Reddy, 2023).

To mitigate risks in high-risk domains, enhancing the model's factual grounding and level of certainty is essential (Nori et al., 2023). Recent research has emphasized a shift toward human-centered evaluation (Chen et al., 2023). Although recent efforts claim that performance improvements stem from encoded high-risk domain knowledge, rendering them applicable in practical real-

world scenarios, certain unexplored directions in evaluation persist. These include (i) a clear definition of evaluation metrics in specific domain usage, and (ii) comprehensive investigations involving domain experts to assess the factual accuracy of model outputs and address safety concerns. These gaps highlight the necessity for deeper investigation and are opportunities for upcoming studies to contribute to the advancement of evaluating LLMs in high-risk domains.

7 Conclusion

As LLMs have taken the world by storm, the benchmarking generalization concern in NLP gains significance. Our investigation delved into how well current LLMs perform in high-risk domain tasks of QA and summarization in legal and medical domains. The results exposed a significant gap of the suitability of LLMs for high-risk domains tasks, indicating that employing LLMs in their present state is *not yet* practical. Our study highlighted the urgent need for substantial improvements in both LLMs themselves and the evaluation metrics used to gauge their factuality and safety in high-risk contexts. Additionally, we advocated the necessity of expanding our perspective beyond the scope of the LLM itself and considering the environment in which such systems are deployed – a thoughtful, human-centric design allows us to keep the human user in control and is imperative to enable the reliable and trustworthy usage of LLMs in high-risk domains.

Overall, our findings and discussions accentuate the importance of a *close* collaboration with stakeholders and therefore *collaboratively* address open critical concerns. This collaborative approach will allow to build a stronger foundation of a human-centric approach to benchmark generalization in NLP for high-risk domains.

Limitations

We investigated how some current LLMs perform on some NLP tasks in the high-risk domains: legal and medical, with regard to two metrics each to

measure factuality and safety. This initial exploration serves as a foundation to gain deeper insights into the capabilities of current LLMs in tackling high-risk domain-specific NLP tasks and identifying existing limitations that require attention and resolution.

The current setup has a series of shortcomings that should be reduced in future work, namely: (1) the collected datasets currently only focus on English; (2) the instruction templates are designed manually and might lead to variable outcomes; (3) other instruction-tuned models trained on general-purpose instructions might offer different capabilities, depending on the specific context of domains and tasks; (4) other metrics should be explored and considered, such as *robustness* (Zhu et al., 2023) and *explainability* (Zhao et al., 2023); and (5) users should be aware that the metrics used are automatic and therefore themselves might also make mistakes and misrepresent model performance (i.e., the metrics require separate benchmarking themselves). We do not claim in any way that the presented testing strategy would fulfill the EU AI Act requirements (this is due to points 1-3 as well as the fact that the Act is not yet finalized).

Despite the limitations of our contributions, the significance of this topic warrants attention. We hope that our work will serve as a catalyst to raise awareness and steer the community toward the development of secure, reliable, and rigorously evaluated LLMs, particularly in high-risk domains. Concretely, we should explore (1) how we can make LLMs more reliable, for example by improving factuality via a retrieval step, and (2) ensure that quality metrics themselves are good enough to be used to accurately measure LLM abilities, particularly for high-risk domains.

Ethics Statement

Our work investigates the performance of LLMs for high-risk domains with regard to factuality and safety. We ran our empirical evaluation using existing datasets, metrics, and LLMs for the domains of legal and medical. At this stage, we did not involve any other stakeholders. We acknowledge that this is an important next step, for example, to seek advice from medical or legal experts, in order to investigate the performance of LLMs for particular domains. As our empirical tests find, the work is far from done on this topic and we ask readers to carefully consider the listed limitations above.

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all. Accessed: 2023-06-03.

Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. Ai chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5:60.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Xiang'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, Chien-Sheng Wu, et al. 2023. Next steps for human-centered generative ai: A technical perspective. *arXiv preprint arXiv:2306.15774*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling

language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

European Commission. 2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206; https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF; https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF. Accessed: 2023-06-22.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. 2023. Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. *arXiv preprint arXiv:2305.07804*.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Laura Hanu and Unitary team. 2020. Detoxify. https://github.com/unitaryai/detoxify. Accessed: 2023-06-15.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Sanjeev Kumar Karn, Rikhiya Ghosh, Kusuma P, and Oladimeji Farri. 2023. shs-nlp at RadSum23: Domain-adaptive pre-training of instruction-tuned LLMs for radiology report impression generation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 550–556, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. 2022. Parameter-efficient legal domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 119–129, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Team MosaicML NLP. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. www.mosaicml.com/blog/mpt-7b. Accessed: 2023-06-03.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. *arXiv preprint arXiv:2306.07075*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2022. chatgpt. https://openai.com/blog/chatgpt. Accessed: 2023-06-23.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, page 101304.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are

continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Ting Fang Tan, Arun James Thirunavukarasu, J Peter Campbell, Pearse A Keane, Louis R Pasquale, Michael D Abramoff, Jayashree Kalpathy-Cramer, Flora Lum, Judy E Kim, Sally L Baxter, et al. 2023. Generative artificial intelligence through chatgpt and other large language models in ophthalmology: Clinical applications and challenges. *Ophthalmology Science*, page 100394.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Annual Symposium*, abs/2008.11293.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax. Accessed: 2023-06-03.

Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D.Q. Bui, Junnan Li, and Steven C. H. Hoi. 2023. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multidimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Examples for Evaluation Tasks

We manually compose the instruction-style templates, designed for each task for evaluation. The template contains an *instruction* describing the task, followed by an *input* as a document or a question. Table 7 shows an example for each evaluation task.

Dataset	Template ‡
BillSum (Kornilova and Eidelman, 2019)	### *Instruction:* Please give a summary of the following legal document: ### *Input:* SECTION 1. TEMPORARY DUTY SUSPENSIONS ON CERTAIN HIV DRUG SUBSTANCES. (a) In General.–Subchapter II of chapter 99 of the Harmonized Tariff Schedule of the United States is amended by inserting in numerical sequence the following new headings: [...] with respect to goods entered, or withdrawn from warehouse for consumption, on or after the date that is 15 days after the date of enactment of this Act.
CaseHold (Zheng et al., 2021)	### *Instruction:* Select one correct answer from ABCDE to match the <HOLDING> statement, not to list all answers. ### *Input:* Statement: has "jurisdiction to render judgment on an action by an interested party objecting [...] A bidder has a direct economic interest if the alleged errors in the procurement caused it to suffer a competitive injury or prejudice. Myers Investigative & Sec. Servs., Inc. v. United States, 275 F.3d 1366, 1370 (Fed.Cir.2002) (<HOLDING>). In a post-award bid protest, the protestor A: holding that an antitrust injury is a necessary element of a 2 claim B: holding that actual prejudice is not a necessary element of an insurers untimely notice defense C: holding that an assertion of prejudice is not a showing of prejudice D: recognizing that allegation of state action is a necessary element of a 1983 claim E: holding that prejudice or injury is a necessary element of standing
LawStackExchange (Li et al., 2022)	### *Instruction:* Please give an answer to the question: ### *Input:* How do we claim the estate of someone who died under a different name in a different country?
PubMedQA (Jin et al., 2019)	### *Instruction:* Answer the question with (yes, no, maybe) and provide the reason based on the given context. ### *Input:* Question: Does oxybutynin hydrochloride cause arrhythmia in children with bladder dysfunction? Context: METHOD: This study represents a subset of a complete data set, considering only those children aged admitted to the Pediatric Surgery and Pediatric Nephrology Clinics during the period January 2011 to July 2012. RESULT: In this study, we have determined that the QT interval changes significantly depending on the use of oxybutynin. The QT changes increased cardiac arrhythmia in children.
RCTSum (Wallace et al., 2020)	### *Instruction:* Summarize the document based on the given title and abstract. ### *Input:* Title: Efficacy of prophylactic antibiotics for the prevention of endomyometritis after forceps delivery. Abstract: The purpose of this prospective randomized controlled clinical trial was to determine whether prophylactic antibiotics reduce the incidence of endomyometritis after forceps delivery. Of the 393 patients studied, 192 received 2 gm of intravenous cefotetan after forceps delivery, and 201 patients received no antibiotics. There were seven cases of endomyometritis in the group given no antibiotic and none in the cefotetan group, a statistically significant difference (P less than .01). We conclude that prophylactic antibiotics are effective in reducing the incidence of endomyometritis after forceps delivery. We believe this is the first published study demonstrating this benefit.
iCliniq (Yunxiang et al., 2023)	### *Instruction:* Please give an answer to the question: ### *Input:* Hello doctor, when should I take probiotics?

Table 7: Templates designed for each evaluation task. ‡For brevity, we record partial inputs for long documents with [...].