Know Your Limits: A Survey of Abstention in Large Language Models

Bingbing Wen[1] **Jihan Yao**[1] **Shangbin Feng**[1] **Chenjun Xu**[1]
Yulia Tsvetkov[1] **Bill Howe**[1] **Lucy Lu Wang**[1,2]
[1]University of Washington, USA [2]Allen Institute for AI, USA
{bingbw, chenjux, billhowe, lucylw}@uw.edu
{jihany2, shangbin, yuliats}@cs.washington.edu

Abstract

Abstention, the refusal of large language models (LLMs) to provide an answer, is increasingly recognized for its potential to mitigate hallucinations and enhance safety in LLM systems. In this survey, we introduce a framework to examine abstention from three perspectives: the query, the model, and human values. We organize the literature on abstention methods, benchmarks, and evaluation metrics using this framework, and discuss merits and limitations of prior work. We further identify and motivate areas for future research, such as whether abstention can be achieved as a meta-capability that transcends specific tasks or domains, and opportunities to optimize abstention abilities in specific contexts. In doing so, we aim to broaden the scope and impact of abstention methodologies in AI systems.[1]

1 Introduction

Large language models (LLMs) have demonstrated generalization capabilities across NLP tasks such as question answering (QA) (Wei et al., 2022a; Chowdhery et al., 2022), abstractive summarization (Zhang et al., 2023a), and dialogue generation (Yi et al., 2024). But these models are also unreliable, having a tendency to ''hallucinate'' false information in their responses (Ji et al., 2023b), generate overly certain or authoritative responses (Zhou et al., 2024b), answer with incomplete information (Zhou et al., 2023b), or produce harmful or dangerous responses (Anwar et al., 2024). In these situations, the model should ideally *abstain*: to refuse to answer in the face of uncertainty (Wen et al., 2024; Feng et al., 2024b; Yang et al., 2023).

Current methods to encourage abstention typically rely on calibration techniques, including linguistic calibration (Mielke et al., 2022; Huang et al., 2024b), which aim to accurately and consistently estimate a model's confidence in its response, then arrange for the model to abstain if the confidence score for a given response falls below some threshold (Varshney et al., 2022; Xiao et al., 2022; Desai and Durrett, 2020). But questions of whether a query is aligned with human values or is answerable at all are difficult to model in terms of model confidence (Yang et al., 2023).

Prior work demonstrates the potential of abstention to enhance model safety and reliability in constrained settings (Varshney et al., 2023; Wang et al., 2024c; Zhang et al., 2024a). In this survey, we attempt to bring together relevant work studying abstention strategies or leading to abstention behaviors, across the diverse range of scenarios encountered by general-purpose chatbots engaging in open-domain interactions. Our goals are to identify gaps and encourage new methods to achieve abstention. Developing or adapting abstention mechanisms to suit a wide array of tasks will enhance the overall robustness and trustworthiness of LLM interactions.

To this end, our survey presents an overview of the current landscape of abstention research. We provide a definition of abstention that incorporates not only technical perspectives—query examination and model capabilities—but also considers alignment with human values. We categorize existing methods to improve abstention in LLMs based on the model lifecycle (pretraining, alignment, and inference), and provide an analysis of evaluation benchmarks and metrics used to assess abstention. In our discussion, we aim to establish a clear entry point for researchers to study the role of abstention across tasks, facilitating the incorporation of new abstention techniques into future LLM systems.

[1]A list of abstention-related papers from this review can be found at https://github.com/chenjux/abstention.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 529–556, 2025. https://doi.org/10.1162/tacl_a_00754
Action Editor: Ellie Pavlick. Submission batch: 8/2024; Revision batch: 12/2025; Published 6/2025.

We summarize our contributions below:

- We introduce a framework to analyze abstention capabilities from three perspectives that have typically been considered in isolation—query answerability, the confidence of the model to answer the query, and alignment of query and responses with human values. Our framework helps us identify existing research that is relevant to abstention as well as abstention mechanisms that have been developed in prior work (§2).
- We conduct a detailed survey of existing abstention methods (§3) as well as evaluation benchmarks and metrics (§4), aiding researchers in selecting appropriate strategies. For each class of methods, we identify opportunities for further research to advance the field.
- We discuss other considerations and underexplored aspects (§5) of abstention, highlighting pitfalls and promising future directions. We encourage researchers to develop more robust model abstention mechanisms and demonstrate their effectiveness in real-world applications.

2 Abstention in LLMs

Definition We define ***abstention*** as the refusal to answer a query. When a model fully abstains, it may begin a response with ‘‘I don’t know’’ or refuse to answer in another way. In reality, abstention encompasses a spectrum of behaviors (Röttger et al., 2024a), e.g., expressing uncertainty, providing conflicting conclusions, or refusing due to potential harm are all forms of abstention. ***Partial abstention*** may involve both answering and abstention, such as self-contradictory responses, e.g., ‘‘I can’t answer the question, but I suppose the answer might be...’’ We do not consider ignoring and/or reframing the question as abstention; but rather as failure modes of LLMs in following instructions (Röttger et al., 2024a; Varshney et al., 2023).

For the ***abstention expression***—the words a model uses to convey that it has abstained—we adopt the definition of five major types of expressions from prior work (Varshney et al., 2023; Wang et al., 2024c), indicating that the model (i) cannot assist; (ii) refutes the query; (iii) provides multiple perspectives without expressing preference; (iv) perceives risk associated with the query and answers cautiously with a disclaimer; and (v) refuses to offer concrete answers due to the lack of knowledge or certainty. Expressions can be identified through heuristic rules and key word matching (Zou et al., 2023; Wen et al., 2024; Yang et al., 2023), or through model-based or human-based evaluation (§4).

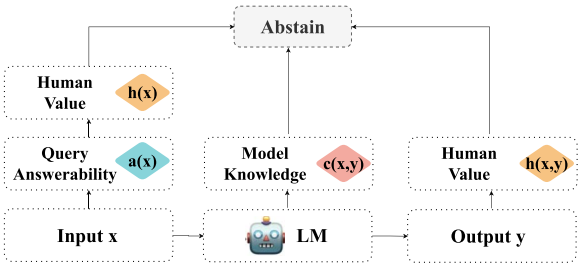

Figure 1: Our proposed framework for abstention in language models. Starting with input query $\mathbf{x}$, the query can be gauged for answerability $a(\mathbf{x})$ and alignment with human values $h(\mathbf{x})$. The model then generates a potential response $\mathbf{y}$ based on the input $\mathbf{x}$. If query conditions are not met, the model’s confidence in the response $c(\mathbf{x}, \mathbf{y})$ is too low, or if the response’s alignment with human values $h(\mathbf{x}, \mathbf{y})$ is too low, the system should abstain.

Below, we describe and motivate our framework for analyzing abstention behavior (§2.1), then provide a formal definition of its components (§2.2).

2.1 Abstention Framework

We study abstention in the scenario of LLMs as AI assistants, exemplified by chatbots such as ChatGPT (OpenAI, 2023; Achiam et al., 2023), Claude (Anthropic, 2023), LLaMA (Touvron et al., 2023), and others (Chiang et al., 2023). We propose an idealized abstention-aware workflow for these systems in Figure 1. Given an LLM f that supports arbitrary generative modeling tasks and the users’ input $\mathbf{x}$, f generates an output $\mathbf{y}$. We analyze the decision to abstain from three distinct but interconnected perspectives:

- The **query** perspective focuses on the nature of the input—whether the query is ambiguous or incomplete (Asai and Choi, 2021), beyond what any human or model could possibly know (Amayuelas et al., 2023), there is irrelevant or insufficient context to answer (Aliannejadi et al., 2019; Li et al., 2024b), or

there are knowledge conflicts (Wang et al., 2023b). In these situations, the system should abstain.

- The **model knowledge** perspective examines the capabilities of the AI model itself, including its design, training, and inherent biases (Ahdritz et al., 2024; Kim and Thorne, 2024; Hestness et al., 2017; Hoffmann et al., 2022; Kaplan et al., 2020; Cao, 2024). For any given query, the system should abstain if the model is insufficiently confident about the correctness of output or has a high probability of returning an incorrect output.
- The **human values** perspective considers ethical implications and societal norms that influence whether a query should be answered, emphasizing the impact of responses on human users (Kirk et al., 2023a). A system should abstain if asked for personal opinions or values (i.e., the query anthropomorphizes the model), or if the query or response may compromise safety, privacy, fairness, or other values.

For examples of queries and outputs meeting conditions for abstention, please see Appendix Table 2.

2.2 Problem Formulation

To formalize our definition of abstention: Consider an LLM $f : \mathcal{X} \to \mathcal{Y}$. When given a prompt $\mathbf{x} \in \mathcal{X}$, f generates a response $\mathbf{y} \in \mathcal{Y}$. We model refusal to answer (abstention) as a function $r : \mathcal{X}, \mathcal{Y} \to [0, 1]$ where $r(\mathbf{x}, \mathbf{y}) = 1$ indicates the system will fully abstain from answering, $r(\mathbf{x}, \mathbf{y}) = 0$ indicates the system will return the output $\mathbf{y}$, and intermediate values represent partial abstention.

We define r as the conjunction of three functions, to be defined by a system designer, to assess query *answerability*, the *confidence* of the LLM's response to the query, and the *human value alignment* of the query and response. We define these three functions as:

- Query function $a : \mathcal{X} \to [0, 1]$. $a(\mathbf{x})$ represents the degree to which an input $\mathbf{x}$ can be answered.
- Model confidence function $c : \mathcal{X}, \mathcal{Y} \to [0, 1]$. $c(\mathbf{x}, \mathbf{y})$ indicates the model f's confidence in its output $\mathbf{y}$ based on input $\mathbf{x}$.
- Human value alignment functions $h : \mathcal{X}, \mathcal{Y} \to [0, 1]$. We define two variants of h: $h(\mathbf{x})$ operates on the input alone and determines its alignment with human values, and $h(\mathbf{x}, \mathbf{y})$ operates on both the input $\mathbf{x}$ and predicted output $\mathbf{y}$. h is measured either through human annotation (Ouyang et al., 2022) or a proxy model that can be learned based on human preferences (Gao et al., 2023).

The refusal function r determines whether the LLM should abstain from responding to input $\mathbf{x}$ as:

$$r(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, \text{if any of } a(\mathbf{x}); c(\mathbf{x}, \mathbf{y}); h(\mathbf{x}, \mathbf{y}) = 0 \\ M(a(\mathbf{x}), c(\mathbf{x}, \mathbf{y}), h(\mathbf{x}, \mathbf{y})), \text{otherwise} \\ 0, \text{if all of } a(\mathbf{x}); c(\mathbf{x}, \mathbf{y}); h(\mathbf{x}, \mathbf{y}) = 1 \end{cases}$$

The function M acts as a connector between the three perspectives, defining how the system integrates their individual outputs into a unified decision. The design of M depends on the system designer and will vary based on the application; examples include weighted averaging, logical operations, or custom thresholds. Some existing systems (Varshney et al., 2022; Cole et al., 2023) use threshold-based mechanisms to convert partial abstention behaviors into binary decisions, such as full compliance or full refusal.

Our framework allows nuanced handling of abstention, by combining confidence from all three perspectives and enabling partial abstention when appropriate. Under this definition, a system would fully abstain from answering if any of the three perspectives indicates full abstention. In all other cases, a system would partially abstain, balancing between providing an answer and withholding information based on indications from the three perspectives.

2.3 Inclusion in this Survey

We identify and survey prior work that falls under any of the three perspectives of our abstention framework. In §3, we organize abstention methodology from an *LLM-centered* perspective, based on when each method is applied in the LLM lifecycle: pretraining, alignment, or inference. This organization is chosen for ease of comparison of experimental settings. Each subsection within §3 is further organized by the three perspectives. Following, §4 describes evaluation benchmarks and metrics that have been used or introduced

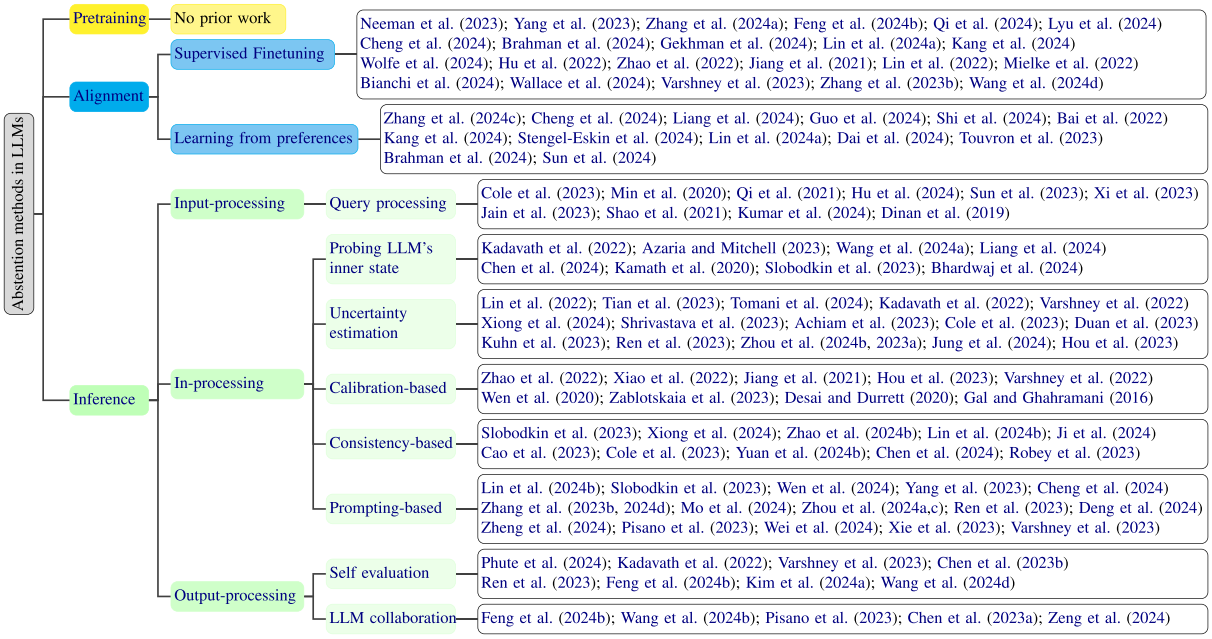

Figure 2: Methods to improve LLM abstention grouped by pretraining, alignment, and inference stages.

across the surveyed prior work. At the end of each subsection, in blue boxes, we summarize main takeaways and provide suggested directions for future work. In §5, we summarize notable threads of research that are not easily classified as method or evaluation.

3 Abstention Methodology

We summarize methods introduced in prior work (Figure 2 organizes these by stages in the LLM lifecycle) and provide ideas for future experiments.

3.1 Pretraining Stage

We found no existing research that studies abstention in the pretraining stage, despite the widely recognized importance of pretraining as a critical phase for model knowledge acquisition. To bridge this gap, we propose several directions for future exploration.

Pretraining Summary

- The impact of refusal-aware data in pretraining is not well-studied.
- Demystify Pretraining Corpora: Investigate how the distribution of refusal-aware data in pretraining corpora impact model abstention ability.
- Coarse-to-fine Abstention Pretraining: In later stages of pretraining, adding data source information including the domain and reasons to abstain into pretraining may be helpful for improving abstention ability.
- Curriculum Learning for Abstention: Simpler abstention scenarios can be introduced first in pretraining, progressing to more complex or fine-grained abstention tasks in later pretraining phases.
- Hybrid Objective Function: Modify the next-word prediction objective in later pretraining stages to incorporate an abstention-aware loss component, balancing between generating tokens and abstaining when appropriate.
- Regularization Techniques: Apply abstention-aware regularization during later stages of pretraining to encourage the model to abstain when prediction uncertainty is high.

3.2 Alignment Stage

We categorize alignment-stage methods as supervised finetuning (SFT) or preference optimization (PO). Some papers include both methods, as PO usually requires SFT as a precursor; we discuss these in the subsection most reflective of their primary contributions.

Supervised Finetuning Many works have demonstrated that SFT with abstention-aware data can improve model abstention capabilities. For example, Neeman et al. (2023) perform data augmentation in the finetuning stage to encourage LLMs to predict unanswerable when presented with an empty or randomly sampled document. Yang et al. (2023) construct an honesty alignment dataset by substituting LLM's wrong or uncertain responses with ''I don't know'' and finetuning on

the resulting data, improving model abstention. Notably, Zhang et al. (2024a) introduce R-tuning, constructing and finetuning on a refusal-aware dataset and showing improved abstention capabilities. Zhang et al. (2024a) also argue that refusal-aware answering is task-independent and could benefit from multi-task training and joint inference. However, Feng et al. (2024b) present contradictory findings in their Appendix that abstention-aware instruction-tuning struggles to generalize across domains and LLMs.

In parallel, concerns have emerged regarding the effectiveness of SFT for abstention. Cheng et al. (2024) and Brahman et al. (2024) find that SFT can make models more conservative, leading to a higher number of incorrect refusals. Recent work (Gekhman et al., 2024; Lin et al., 2024a; Kang et al., 2024) also demonstrates that finetuning on examples unobserved during pretraining increases the risk of hallucination. Gekhman et al. (2024) propose a mitigation strategy to re-label these examples based on the pretrained LLM's knowledge and include ''I don't know'' in the finetuning data to teach the model to abstain. A related method to reduce hallucination is introduced in Lin et al. (2024a); the authors create factuality-aware training data for SFT by classifying whether an instruction requires a factual response.

Parameter-efficient finetuning (PEFT) strategies have also been used for abstention. Wolfe et al. (2024) conduct lab-scale experiments, finetuning LLMs with QLoRA (Dettmers et al., 2023), and observe that weaker models (with lower task performance) tend to achieve greater gains in abstention performance. Beyond resource efficiency, Brahman et al. (2024) have found that LoRA (Hu et al., 2022) acts as an effective regularization technique for improving abstention; they find that fully finetuned models exhibit over-refusal while also forgetting general capabilities, and demonstrate that finetuning with LoRA alleviates both issues while significantly improving abstention behavior.

Instead of finetuning directly, finetuning for calibration may indirectly improve abstention ability (Szegedy et al., 2016; Zhao et al., 2022; Xiao et al., 2022; Jiang et al., 2021; Lin et al., 2022). Jiang et al. (2021) propose two finetuning objective functions (softmax-based and margin-based), which improve Estimated Calibration Error (ECE) (Guo et al., 2017a) on multiple-choice datasets. Mielke et al. (2022) alternatively use a calibrator trained to provide a confidence score with an LLM finetuned to control for linguistic confidence in a system.

Towards alignment with human values, Bianchi et al. (2024) show that adding a small number of safety instructions to instruction-tuning data reduces harmful responses without diminishing general capabilities, whereas an excessive number of safety instructions makes LLMs overly defensive. Varshney et al. (2023) construct responses for unsafe prompts by combining fixed refusal responses with Llama-2-generated safe responses, and obtain similar results. Wallace et al. (2024) finetune LLMs to follow hierarchical prompts, enhancing the fine-grained abstention ability of LLMs. Zhang et al. (2023b) also finetune LLMs with goal prioritization instructions that instruct LLMs to prioritize safety over helpfulness during inference.

However, custom finetuning of LLMs presents safety risks. For example, Qi et al. (2024); Lyu et al. (2024) note that finetuning with benign and commonly used datasets increases unsafe behaviors in aligned LLMs (Qi et al., 2024). To address this, Lyu et al. (2024) propose to finetune models without a safety prompt, but include one at test time. Wang et al. (2024d) finetune LLMs to evaluate their own outputs for harm and append a ''harmful'' or ''harmless'' tag to its responses instead of directly tuning LLMs to abstain.

Instruction Tuning Summary

- Instruction tuning on abstention-aware data improves abstention ability but can lead to over-abstention.
- Researchers disagree on whether instruction tuning helps LLMs learn abstention as a meta-capability.

Heterogeneous approaches: Abstention-aware instruction-tuning datasets tend to emphasize only a few question-response formats. Incorporating diverse abstention expressions, definitions, and domains may improve generalization. Ensembling models trained in different abstention domains, e.g., mixture of LoRA experts, may also help achieve better generalization.

Learning from Preferences Preference optimization can impact abstention from both the model knowledge and human value alignment perspectives. As described above, finetuning LLMs on abstention-aware data may lead to overly conservative behavior, causing erroneous refusals of queries. Cheng et al. (2024) and Brahman et al. (2024) address this through Direct Preference Optimization (DPO) (Rafailov et al., 2023), encouraging the model to answer questions it knows and refuse questions it does not know.

Factuality-based preference optimization can help models respond correctly to queries, including abstaining (e.g., saying ‘‘I don’t know’’). As an example, Liang et al. (2024) construct a factual preference dataset to train a reward model, and utilize it to optimize abstention preferences in LLMs via Proximal Policy Optimisation (PPO) (Schulman et al., 2017). Kang et al. (2024) design a reward function that prioritizes abstention over incorrect answers, while Lin et al. (2024a) incorporate factuality-focused preference pairs into DPO to enhance fact-based instruction following.

Other works use DPO to improve calibration, which can also aid abstention. LACIE (Stengel-Eskin et al., 2024) casts confidence calibration as a preference optimization problem and introduce a speaker-listener game to create preference data; they demonstrate that finetuning on LACIE data leads to emergent model abstention behavior. Zhang et al. (2024c) introduce Self-Alignment for Factuality, generating confidence scores through self-asking to improve calibration via DPO.

For human values, safety alignment methods (Dai et al., 2024; Touvron et al., 2023; Bai et al., 2022; Shi et al., 2024) use explicit or implicit preference models to reduce harmfulness, which though not explicitly focused on abstention, will encourage abstention on unsafe prompts. Other studies have explored multi-objective alignment approaches (Guo et al., 2024) to encourage safe and helpful model behavior. The instructable reward model in SALMON (Sun et al., 2024) is trained on synthetic preference data, generating reward scores based on customized human-defined principles as the preference guideline.

Learning from Preferences Summary

- Preference optimization can reduce over-abstention caused by instruction tuning on refusal-aware data by aligning model behavior with user preferences. However, this may not always be the case. Preference optimization may still lead to over-abstention if reward models overemphasize safety or preference data favors abstention. Balancing feedback and ensuring diverse preference data are crucial to address this challenge, making it an important area for further research.

💡 Ranking-based preference optimization: Since abstention is a spectrum of behaviors, ranking-based preference optimization can extend the pair-wise contrast to accommodate rankings over different degrees of abstention.

3.3 Inference Stage

We categorize inference stage methods as input-processing, in-processing, or output-processing approaches based on when they are applied. Input-processing approaches are centered on the query answerability and human values perspectives; in-processing approaches on the model knowledge perspective; and output-processing approaches may consider both model knowledge and human values.

3.3.1 Input-processing Approaches

Query Processing From the query perspective in our proposed framework, LLMs can choose to abstain based on the query answerability. For example, Cole et al. (2023) try to predict the ambiguity of questions derived from the AmbigQA dataset (Min et al., 2020) before selectively answering.

Other methods aim to identify queries that are misaligned with human values. For example, Qi et al. (2021) detect malicious queries needing abstention by removing suspect words from the query and analyzing the resulting drop in perplexity while Hu et al. (2024) propose new ways of computing perplexity and find tokens with abnormally high perplexity. Apart from perplexity-based methods, Jain et al. (2023) further investigate input preprocessing methods such as paraphrasing and retokenization. The BDDR framework (Shao et al., 2021) not only detects suspicious words in the input but also reconstructs the original text through token deletion or replacement. Kumar et al. (2024) introduce the ‘‘erase-and-check’’ framework to defend against adversarial prompts with certifiable safety guarantees. Similarly, Xi et al. (2023) measure changes in representation between original and paraphrased queries using a set of distributional anchors to identify harmful queries. Dinan et al. (2019) develop a more robust offensive language detection system through an iterative build-it, break-it, fix-it strategy.

Query Processing Summary

- Query processing approaches focus on assessing query ambiguity or human values alignment.

💡 Context awareness: Existing work focuses on query-only processing or simple context-dependent tasks insufficient or conflicting context may be provided in the abstention scenario (Rajpurkar et al., 2018; Kwiatkowski et al., 2019). These settings overlook

context complexity in real-world applications; for example, earlier context in multi-turn conversations can impact judgments for either query answerability or human values alignment in later conversational turns.

3.3.2 In-processing Approaches

Probing LLM's Inner State Recent studies (Kamath et al., 2020; Azaria and Mitchell, 2023) focus on training calibrators based on LLM internal representation to predict the accuracy of the model's responses, enabling abstention when the likelihood of error is high. Further probing into the internal representations of LLMs to discern between answerable and unanswerable queries has been conducted by Slobodkin et al. (2023), Kadavath et al. (2022) and Liang et al. (2024). Additionally, Chen et al. (2024) introduce the EigenScore, a novel metric derived from LLM's internal states, which can facilitate abstention by quantifying the reliability of the model's knowledge state.

In terms of leveraging the LLMs' internal states for safety judgments, Wang et al. (2024a) extract safety-related vectors (SRVs) from safety-aligned LLMs; which are then used as an abstention gate to steer unaligned LLMs towards safer task performance. Furthermore, Bhardwaj et al. (2024) demonstrate that integrating a safety vector into the weights of a finetuned LLM through a simple arithmetic operation can significantly mitigate the potential harmfulness of the model's responses.

Probing LLM's Inner State Summary

- LLM's internal representations can indicate model knowledge boundaries and safety awareness.

Uncertainty Estimation Estimating the uncertainty of LLM output can serve as a proxy for making abstention decisions. Token-likelihoods have been widely used to assess the uncertainty of LLM responses (Lin et al., 2022; Kadavath et al., 2022). Enhancing this approach, Lin et al. (2022) and Tian et al. (2023) employ an indirect logit methodology to calculate the log probability of the 'True' token when appended to model's generated response. Shrivastava et al. (2023) leverage a surrogate LLM with access to internal probabilities to approximate the confidence of the original model. Tomani et al. (2024) assess Predictive Entropy and Semantic Entropy (Kuhn et al., 2023) of responses. Duan et al. (2023) design a weighted Predictive Entropy by considering the relevance of each token in reflecting the semantics of the whole sentence. However, other work shows that aligned LLMs may not have well-calibrated logits (Cole et al., 2023; Achiam et al., 2023) and may have positional bias and probability dispersion (Ren et al., 2023). In the context of LLM-as-judge, these canonical probability-based methods tend to be overconfident in estimating agreement with the majority of annotators; Jung et al. (2024) propose a novel confidence estimation method by simulating diverse annotator preferences with in-context learning.

The Maximum Softmax Probability approach (Varshney et al., 2022) uses peak softmax output as a uncertainty estimator. Hou et al. (2023) introduce an uncertainty estimation method: input clarification ensembling. Through ruling out aleatoric uncertainty by clarification, the remaining uncertainty of each individual prediction is epistemic uncertainty.

Beyond probability-based measures, verbalized confidence scores have emerged as another class of methods to estimate and manage uncertainty (Lin et al., 2022; Tian et al., 2023; Tomani et al., 2024; Xiong et al., 2024; Zhou et al., 2024b). Xiong et al. (2024) examine prompting methods including chain-of-thought (Wei et al., 2022b), self-probing, top-k (Tian et al., 2023), and linguistic likelihood expressions to eliciting confidence scores. Although LMs can be explicitly prompted to express confidence, verbalized confidence scores have been found to be over-confident (Xiong et al., 2024; Zhou et al., 2024b). Zhou et al. (2024b) find that LMs are reluctant to express uncertainty when answering questions, even when their responses are incorrect. Zhou et al. (2023a) show that high-certainty expressions in the prefix of a response can result in accuracy drop compared to low-certainty expressions, suggesting that LLMs respond more to prompting style rather than accurately assessing epistemic uncertainty.

Uncertainty Estimation Summary

- Uncertainty can act as proxy for abstention. However, neither probability-based measures nor verbalized confidence may be well-calibrated.
- Uncertainty estimation could be used as a key aspect of explainable AI. The model could communicate its uncertainty or insufficient confidence in its response, which could foster user trust and engagement by making the model's decision-making process more transparent.

Calibration-based Methods Estimated model uncertainty may not accurately represent the likelihood of a model's outputs being correct, so numerous studies focus on calibrating the uncertainty of LLMs. Jiang et al. (2021) improve calibration by augmenting inputs and paraphrasing outputs. Temperature Scaling (Guo et al., 2017b; Xiao et al., 2022; Desai and Durrett, 2020; Jiang et al., 2021) modifies the softmax temperature to refine calibration during decoding. Additionally, Monte-Carlo Dropout (Gal and Ghahramani, 2016; Varshney et al., 2022; Zablotskaia et al., 2023) employs multiple predictions with varying dropout configurations to assemble a robust confidence estimate. Batch Ensemble (Wen et al., 2020) is a computationally efficient method that aggregates multiple model predictions and maintains good calibration.

Calibration-based Methods Summary

- Confidence calibration is a longstanding area of research; methods developed for calibration can also enhance abstention capabilities, improving model output reliability.
- Domain disparity: Investigating methods to ensure calibration techniques remain effective across diverse datasets and different types of models, particularly under domain shifts, can surface task-specific optimizations.

Consistency-based Methods Given the limitations of confidence elicitation, some methods leverage consistency-based aggregation to estimate LLM uncertainty and then abstain when uncertain. Aggregation can be achieved using diversity and repetition (Cole et al., 2023), weighted confidence scores and pairwise ranking (Xiong et al., 2024), or semantic similarity between responses (Lin et al., 2024b; Zhao et al., 2024b; Chen et al., 2024). Slobodkin et al. (2023) relax beam search and abstain if any top-k answer is ''unanswerable''.

Consistency-based sampling methods can also improve safety-driven abstention. Robey et al. (2023), Cao et al. (2023), and Ji et al. (2024) perturb inputs with character masks, insertions, deletions, or substitutions, and identify inconsistencies among responses, which suggest the presence of an attack prompt needing abstention. Yuan et al. (2024b) obtain samples by prompting for augmentations (learnable safe suffixes and paraphrasing) and use a kNN-based method to aggregate responses.

Consistency-based Methods Summary

- Consistency-based methods establish model certainty based on its output distribution, helping to identify queries for which a model should abstain.

Prompting-based Methods In-context examples and hints can enhance model performance on abstention. Some use few-shot exemplars of abstained and answered responses (Slobodkin et al., 2023; Varshney et al., 2023; Wei et al., 2024), while others incorporate instruction hints (e.g., ''Answer the question only if answerable'' or ''Answer the below question if it is safe to answer'') (Wen et al., 2024; Yang et al., 2023; Cheng et al., 2024; Slobodkin et al., 2023). For multiple-choice QA, adding ''None of the above'' as an answer option has been shown to be effective (Ren et al., 2023; Lin et al., 2024b). Zhang et al. (2023b) explicitly prompt LLMs to prioritize safety over helpfulness. Deng et al. (2024) also propose that providing explanations on the unanswerability of questions not only improves model explainability, but can produce more accurate responses.

Other work focuses on carefully designed prompts. Mo et al. (2024) concatenate a protective prefix from attack-defense interactive training with the user query. Similarly, Zhou et al. (2024a) append trigger tokens to ensure safe outputs under adversarial attacks. Pisano et al. (2023) use another LLM to add conscience suggestions to the prompt. Zhang et al. (2024d) prompt LLMs to analyze input intent and abstain if malicious. Xie et al. (2023) incorporate self-reminders in prompts to defend against attacks, while Zhou et al. (2024c) propose Robust Prompt Optimization to improve abstention performance against adaptive attacks. Zheng et al. (2024) find that safety prompts can safeguard LLMs against harmful queries and further propose a safety prompt optimization method to shift query representations toward or away from the refusal direction based on query harmfulness.

Prompting-based Methods Summary

- Prompting with abstention examples demonstrates potential in enhancing the abstention capabilities of LLMs.
- Interpretable prompting methods: Developing methods to effectively use instructions or choose demonstrations that capture various abstention behaviors remains under-explored, along with identifying the limits of what can be achieved through in-context learning.

3.3.3 Output-processing Approaches

Self Evaluation Chen et al. (2023b) use Soft Prompt Tuning to learn self-evaluation parameters for various tasks. However, directly asking LLMs to evaluate if their responses are certain or safe (usually in a different conversation), and to abstain if they are not, has proven effective in improving LLM abstention (Phute et al., 2024; Kadavath et al., 2022; Varshney et al., 2023; Ren et al., 2023; Feng et al., 2024b). Kim et al. (2024a) allow the LLM to iteratively provide feedback on its own responses and refine its answers; this method achieves improvements in safety even in non-safety-aligned LLMs. Wang et al. (2024d) enable LLMs to self-evaluate responses and append a [harmful] or [harmless] tag to each response; however, this approach may encourage over-abstention.

LLM Collaboration Multi-LLM systems are effective in producing better overall responses, including improved abstention behavior. In 2-LLM systems, a test LLM is employed to examine the output of the first LLM and helps with abstaining. In Wang et al. (2024b), the test LLM is used to guess the most likely harmful query from the output and abstains if a harmful query is detected. Pisano et al. (2023) critique and correct a model's original compliant response using a secondary LLM.

Multi-LLM systems beyond two LLMs leverage different LLMs as experts to compete or cooperate to reach a final abstention decision (Feng et al., 2024b; Chen et al., 2023a). As an example, Zeng et al. (2024) employ a group of LLMs in a system with an intention analyzer, original prompt analyzer, and judge.

Self Evaluation & LLM Collaboration Summary

- LLMs can struggle with self-evaluation, but combining multiple LLMs has been effective in enhancing abstention capabilities.

4 Evaluation of Abstention

We survey evaluation benchmarks (§4.1) and metrics (§4.2) used to assess abstention capabilities.

4.1 Evaluation Benchmarks

Below, we describe benchmarks that include abstention in their ground truth annotations; additional dataset details are provided in Appendix Table 3. Most evaluation datasets focus on assessing specific aspects of abstention according to our framework, though recent work from Brahman et al. (2024) espouse a holistic evaluation strategy.

Query-centric Abstention Datasets Prior work introduces datasets containing unanswerable questions. SQuAD2 (Rajpurkar et al., 2018) first includes unanswerable questions with *irrelevant* context passages for machine reading comprehension. Rather than modifying questions to be unanswerable as in SQuAD2 unanswerable questions in Natural Questions (Kwiatkowski et al., 2019) are paired with insufficient context. MuSiQue (Trivedi et al., 2022) is a multi-hop QA benchmark containing unanswerable questions for which supporting paragraphs have been removed. CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) introduce unanswerable questions for conversational QA. Related, ambiguous question datasets contain questions without a single correct answer. AmbigQA (Min et al., 2020) extracts questions from NQ-Open (Kwiatkowski et al., 2019) with multiple possible answers. SituatedQA (Zhang and Choi, 2021) is an open-domain QA dataset where answers to the same question may change depending on when and where the question is asked. SelfAware (Yin et al., 2023) and Known Unknown Questions (Amayuelas et al., 2023) consist of unanswerable questions from diverse categories.

Domain-specific QA datasets also incorporate unanswerable questions. PubmedQA (Jin et al., 2019) contains biomedical questions that can be answered ‘‘yes’’, ‘‘no’’, or ‘‘maybe’’; where ‘‘maybe’’ indicates *high uncertainty* based on the given context. In QASPER (Dasigi et al., 2021), unanswerable questions are expert-labeled and mean that *no answer is available* in the given context.

Model Knowledge-centric Abstention Datasets RealTimeQA (Kasai et al., 2023) is a dynamic dataset which announces questions and evaluates systems on a regular basis, and contains inquiries about current events. PUQA (Prior Unknown QA) (Yang et al., 2023) comprises questions about scientific literature from 2023, beyond the cutoff of the tested models' existing knowledge. ElectionQA23 (Feng et al., 2024b) is a QA dataset focusing on 2023 elections around the globe; due to the temporality of training data, LLMs lack

up-to-date information to accurately respond to these queries. Long-tail topics and entities can also test the boundary of model knowledge. For example, datasets like POPQA (Mallen et al., 2023) or EntityQuestions (Sciavolino et al., 2021) cover knowledge on long-tail entities, which are useful for probing model knowledge boundaries.

Human Value-centric Abstention Datasets Here are datasets designed to measure whether LLM outputs are ‘‘safe,’’ i.e., align with widely held ethical values; these datasets may consist of prompts that are either inherently unsafe or likely to elicit unsafe responses from LLMs. Some datasets focus on specific aspects of safety. A main concern is toxicity, when models generate harmful, offensive, or inappropriate content. For instance, RealToxicityPrompts (Gehman et al., 2020) gathers prompts to study toxic language generation, while ToxiGen (Hartvigsen et al., 2022) and LatentHatred (ElSherief et al., 2021) address implicit toxic speech, and ToxicChat (Lin et al., 2023) collects data from real-world user–AI interactions. Beyond toxicity, Beavertails (Ji et al., 2023a) balances safety and helpfulness in QA, CValues (Xu et al., 2023a) assesses safety and responsibility, and Xstest (Röttger et al., 2024a) examines exaggerated safety behaviors. Latent-Jailbreak (Qiu et al., 2023) introduces a benchmark that assesses both the safety and robustness of LLMs. Do-Anything-Now (Shen et al., 2023) collects a set of unsafe prompts for malicious purposes.

Comprehensive safety benchmarks attempt to encompass a range of concerns. Röttger et al. (2024b) conduct the first systematic review of open datasets for evaluating LLM safety. Do-Not-Answer (Wang et al., 2024c) includes instructions covering information hazards, malicious uses, discrimination, exclusion and toxicity, misinformation harms, and human-computer interaction harms. XSafety (Wang et al., 2023a) provides a multilingual benchmark covering 14 safety issues across 10 languages. SALAD-Bench (Li et al., 2024a) is a large-scale dataset with a three-tier taxonomy, evaluating LLM safety and attack-defense methods. SORRY-Bench (Xie et al., 2024) proposes a more fine-grained taxonomy and diverse instructions. Most relevant to abstention, WildGuard (Han et al., 2024) evaluates model refusal performance as a necessary component for safety.

GT \ R	Correctly answered	Incorrectly answered	Abstained
No abstention	N_1	N_2	N_3
Abstention		N_4	N_5

Table 1: Abstention confusion matrix. ‘‘GT’’: ground-truth human label, where ‘‘Abstention’’ indicates questions labeled as those where the model should abstain. ‘‘R’’: system response. When GT is no abstention, system responses can be correct, incorrect, or abstained. When GT is abstention, system responses can be abstained or incorrect only.

Evaluation Benchmarks Summary

Existing benchmarks primarily focus on a single perspective. Researchers could benefit from more comprehensive benchmarks encompassing examples across the query, model, and human values perspectives, that are capable of system-wide assessment.

4.2 Evaluation Metrics

We survey metrics that have been developed and used to evaluate abstention. Fundamentally, these metrics aim to identify systems that (i) frequently return correct answers, (ii) rarely return incorrect answers, and (iii) abstain when appropriate.

Statistical Automated Evaluation We express these metrics based on the abstention confusion matrix in Table 1.

- *Abstention Accuracy (ACC)* (Feng et al., 2024b) evaluates the system’s overall performance when incorporating abstention:

$$\text{ACC} = \frac{N_1 + N_5}{N_1 + N_2 + N_3 + N_4 + N_5}$$

- *Abstention Precision* (Feng et al., 2024b) measures the proportion of model abstain decisions that are correct:

$$\text{Precision}_{abs} = \frac{N_5}{N_3 + N_5}$$

- *Abstention Recall* (Feng et al., 2024b; Cao et al., 2023; Varshney et al., 2023) or *Prudence Score* (Yang et al., 2023) measures the proportion of cases where models correctly abstain when they should:

$$\text{Recall}_{abs} = \frac{N_5}{N_2 + N_4 + N_5}$$

- *Attack Success Rate* or *Unsafe Responses on Unsafe Prompts (URUP)* (Cao et al., 2023; Varshney et al., 2023) reports the proportion of cases where models do not abstain when they should (indicating successful attacks):

$$\text{URUP} = 1 - \text{Recall}_{abs}$$

- *Abstention F1-score* (Feng et al., 2024b) combines abstention precision and recall:

$$\text{F1}_{abs} = 2 \cdot \frac{\text{Precision}_{abs} \cdot \text{Recall}_{abs}}{\text{Precision}_{abs} + \text{Recall}_{abs}}$$

- *Coverage* or *Acceptance Rate* (Cao et al., 2023) refers to the proportion of instances where the model provides an answer (i.e., does not abstain); it measures the model's willingness to respond:

$$\text{Coverage} = \frac{N_1 + N_2 + N_4}{N_1 + N_2 + N_3 + N_4 + N_5}$$

- *Abstention Rate* (Wen et al., 2024; Varshney et al., 2023), on the other hand, measures the proportion of queries where the model abstains:

$$\text{Abstention Rate} = \frac{N_3 + N_5}{N_1 + N_2 + N_3 + N_4 + N_5}$$

- *Benign Answering Rate (BAR)* (Cao et al., 2023) focuses only on queries deemed to be safe:

$$\text{BAR} = \frac{N_1 + N_2}{N_1 + N_2 + N_3}$$

- *Over-conservativeness Score* or *Abstained Responses on Safe Prompts (ARSP)* (Yang et al., 2023; Varshney et al., 2023) computes the proportion of queries where the model over-abstains:

$$\text{ARSP} = \frac{N_3}{N_1 + N_2 + N_3}$$

- *Reliable Accuracy (R-Acc)* (Feng et al., 2024b) indicates to what extent LLM-generated answers can be trusted when they *do not* abstain, i.e., of all questions answered, how many are correct:

$$\text{R-Acc} = \frac{N_1}{N_1 + N_2 + N_4}$$

- *Effective Reliability (ER)* (Feng et al., 2024b; Si et al., 2023; Whitehead et al., 2022) strikes a balance between reliability and coverage, i.e., of all questions, how many more are answered correctly than incorrectly:

$$\text{ER} = \frac{N_1 - N_2 - N_4}{N_1 + N_2 + N_3 + N_4 + N_5}$$

- *Abstain Estimated Calibration Error (Abstain ECE)* (Feng et al., 2024b) modifies traditional ECE (Guo et al., 2017a) by including abstention. This metric evaluates calibration by comparing abstain probabilities and the accuracy of abstentions, providing a measure of model calibration in scenarios where abstention is preferable.

- *Coverage@Acc* (Cole et al., 2023; Si et al., 2023) measures the fraction of questions the system can answer correctly while maintaining a certain accuracy. Specifically, C@Acc is the maximum coverage such that the accuracy on the C% of most-confident predictions is at least Acc%.

- *Area Under Risk-Coverage Curve (AURCC)* (Si et al., 2023; Yoshikawa and Okazaki, 2023) computes, for any given threshold, an associated coverage and error rate (risk), which is averaged over all thresholds. Lower AURCC indicates better selective QA performance.

- *Area Under Accuracy-Coverage Curve (AUACC)* (Cole et al., 2023; Xin et al., 2021) computes, for any given threshold, an associated coverage and accuracy, which is averaged over all thresholds. Higher AUACC indicates better performance.

- *Area Under Receiver Operating Characteristic curve (AUROC)* (Cole et al., 2023; Kuhn et al., 2023) evaluates the uncertainty estimate's diagnostic ability as a binary classifier for correct predictions by integrating over the tradeoff curve between rates of true and false positives.

Statistical Automated Evaluation Summary

- **Overall performance**: No single metric captures all aspects of performance. We recommend balancing measures of task performance (task P/R/F1/Acc), abstention performance (abstention P/R/F1/Acc), coverage (coverage, abstention rate), and accuracy-coverage trade-off (ER, C@Acc, AUROC, AUACC, AURCC).
- **Error rates**: To assess abstention-related error rates, URUP measures the false answering rate and ARSP the false abstention rate.

Evaluating partial abstention Existing statistical automated evaluation methods focus on identifying and evaluating full abstention, but partial abstention can be incorporated as a weighted sum of whether the response includes abstention and/or answers. By weighting abstention and answer accuracy, we can better qualify a model's actual performance.

Model-based Evaluation Many studies implement LLM-as-a-judge for abstention evaluation (Mazeika et al., 2024; Souly et al., 2024; Chao et al., 2024). Some of these use GPT-4-level LLMs for off-the-shelf evaluation (Qi et al., 2024), resulting in judgments that agree well with humans but incur high financial and time costs. Others explore supplementary techniques to boost the accuracy of the LLM judge such as (i) Chain-of-thought prompting: asking the LLM to ''think step-by-step'' before deciding whether to not answer (Qi et al., 2024; Xie et al., 2024); (ii) In-context-learning: using refusal annotations from a training set as in-context examples (Xie et al., 2024); or (iii) Finetuning LLMs for abstention evaluation (Huang et al., 2024a; Li et al., 2024a). Röttger et al. (2024a) extended full abstention evaluation by prompting GPT-4 with a taxonomy to classify responses as full compliance, full refusal, or partial refusal in a zero-shot setting.

Model-based Evaluation Summary

Model-based evaluations focus on the human values perspective, particularly safety. Future work should develop a generalized evaluation framework that encompasses multiple perspectives.

Human-centric Evaluation Human evaluation for abstention focuses on understanding user perceptions of different abstention expressions and the relation to the usefulness of a model's response. Instead of binary decisions (full compliance and full refusal), Röttger et al. (2024a) introduce partial refusal when manually annotating model's response. Wester et al. (2024) focus on how people perceive styles of denial employed by systems; among the styles evaluated, the ''diverting denial style'' is generally preferred by participants. Kim et al. (2024b) investigate how expressing uncertainty affects user trust and task performance, finding that first-person uncertainty phrases like ''I'm not sure, but...'' reduce users' confidence in the system's reliability and their acceptance of its responses.

Human-centric Evaluation Summary

Human evaluation methods focus on full abstention and strong abstention expression categories, overlooking the significance of partial abstention and other forms of abstention expressions. Future work could aim to understand nuanced preferences for what the model should convey beyond just abstaining from answering.

5 Other Considerations for Abstention

Over-abstention Over-abstention occurs when models abstain unnecessarily. For example, Varshney et al. (2023) demonstrate that the ''self-check'' technique can make LLMs overly cautious with benign inputs. Others similarly observe that instruction tuning with excessive focus on abstention can lead models to inappropriately refuse to respond (Cheng et al., 2024; Bianchi et al., 2024; Wallace et al., 2024; Brahman et al., 2024). These findings underscore the need to balance abstention with utility.

Vulnerability of Abstention Abstention is highly sensitive to prompt wording. Safety-driven abstention mechanisms are notably susceptible to manipulation. Studies show that social engineering techniques such as persuasive language and strategic prompt engineering can bypass established safety protocols (Xu et al., 2023b; Chao et al., 2023). Even ostensibly benign approaches like finetuning with safe datasets or modifying decoding algorithms can inadvertently undermine the safety alignment of LLMs (Qi et al., 2024; Huang et al., 2024a). Advanced manipulation tactics include persona-based attacks (Shah et al., 2023), cipher-based communications (Yuan et al., 2024a), and the translation of inputs into low-resource languages (Yong et al., 2023; Feng et al., 2024a). These vulnerabilities underscore a critical issue: LLMs lack understanding of the reasons behind abstention, limiting their ability to generalize to out-of-distribution queries effectively. Furthermore, objectives like helpfulness and abstention may conflict, and models

may struggle to abstain appropriately in situations where they are confident in their ability to provide helpful responses.

Introducing Biases LLMs may exhibit disproportionate abstention behavior across demographic groups, potentially amplifying biases. For example, Xu et al. (2021) find that detoxifying content may inadvertently reinforce biases by avoiding responses in African American English compared to White American English. Feng et al. (2024b) show that LLMs abstain less when predicting future election outcomes for Africa and Asia in ElectionQA23, raising fairness concerns as these mechanisms might underserve marginalized communities and countries. More work is needed to clarify and address these performance disparaties.

Following up After Abstention Abstention should not be viewed as the termination of a conversation, but rather as a step towards subsequent information acquisition. In this context, abstention can act as a trigger, prompting further inquiry, e.g., asking the user for more information or retrieving additional relevant data (Feng et al., 2024b; Li et al., 2024b). After abstaining, systems should seek out more information when appropriate, transforming abstention from a static endpoint into a dynamic, constructive component of dialogue progression. For example, Zhao et al. (2024a) study the alternate task of reformulating unanswerable questions to questions that can be answered by a given document.

Personalized Abstention Users have different preferences for model abstention (Wester et al., 2024) based on individual differences (Zhang et al., 2024b) and task-specific needs, and no one-size-fits-all solution exists (Kirk et al., 2023b). Personalized abstention mechanisms in LLMs will allow the model to dynamically adjust its abstention behavior based on a user's profile, tolerance for conservative responses, interaction history, specific query needs, and any other requirements.

6 Future Directions

There are many under-explored and promising research directions in abstention, some of which are described in this survey. While prior work has explicitly investigated abstention in specific tasks or implicitly contributed to improved abstention behaviors, we encourage study of abstention as a meta-capability across tasks, as well as more generalizable evaluation and customization of abstention capabilities to user needs. Beyond what has been discussed previously, other important directions include: (i) enhancing privacy and copyright protections through abstention-aware designs to prevent the extraction of personal private information and copyrighted text fragments; (ii) generalizing the concept of abstention beyond LLMs to vision, vision-language, and generative machine learning applications; and (ii) improving multilingual abstention, as significant performance discrepancies exist between high-resource and low-resource languages, necessitating further research to ensure consistent performance across different languages.

7 Conclusion

Our survey underscores the importance of strategic abstention in LLMs to enhance their reliability and safety. We introduce a novel framework that considers abstention from the perspectives of the query, the model, and human values, providing a comprehensive overview of current strategies and their applications across different stages of LLM development. Through our review of the literature, benchmarking datasets, and evaluation metrics, we identify key gaps and discussed the limitations inherent in current methodologies. Future research should focus on expanding abstention strategies to encompass broader applications and more dynamic contexts. By refining abstention mechanisms to be more adaptive and context-aware, we can further the development of AI systems that are not only more robust, reliable, and aligned with ethical standards and human values, but balance these goals more appropriately against helpfulness to the user.

Acknowledgments

This research was supported in part by the National Science Foundation under CAREER Grant No. IIS2142739, and by the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We also gratefully acknowledge support from the UW iSchool Strategic Research Fund and the

University of Washington Population Health Initiative, as well as gift funds from the Allen Institute for AI. We thank the authors of the cited papers for reviewing our descriptions of their work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM Sigir Conference on Research and Development in Information Retrieval*, pages 475–484. https://doi.org/10.1145/3331184.3331265

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*. https://doi.org/10.18653/v1/2024.findings-acl.383

Anthropic. 2023. Introducing claude.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó. hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.118

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.findings-emnlp.68

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. Language models are Homer Simpson! Safety re-alignment of fine-tuned language models through task arithmetic. https://doi.org/10.18653/v1/2024.acl-long.762

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2023. Defending against alignment-breaking attacks via robustly aligned LLM. *arXiv preprint arXiv:2309.14348*.

Lang Cao. 2024. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.212

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Bocheng Chen, Advait Paliwal, and Qiben Yan. 2023a. Jailbreaker in jail: Moving target defense for large language models. In *Proceedings of the 10th ACM Workshop on Moving Target Defense*, pages 29–32. https://doi.org/10.1145/3605760.3623764

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023b. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213. Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.345

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can AI assistants know what they don't know? In *Forty-first International Conference on Machine Learning*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1241

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*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.365

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't just say ''I don't know''! self-aligning large language models for responding to unknown questions with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.757

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.21

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1461

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya

Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*. https://doi.org/10.18653/v1/2024.acl-long.276

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.29

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024a. Teaching LLMs to abstain across languages via multilingual feedback. *arXiv preprint arXiv:2406.15948*. https://doi.org/10.18653/v1/2024.emnlp-main.239

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying LLM knowledge gaps via Multi-LLM collaboration. *arXiv preprint arXiv:2402.00367*. https://doi.org/10.18653/v1/2024.acl-long.786

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*. https://doi.org/10.18653/v1/2024.emnlp-main.444

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017a. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.234

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang

Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv: 2203.15556*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*.

Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhengmian Hu, Gang Wu, Saayan Mitra, Ruiyi Zhang, Tong Sun, Heng Huang, and Viswanathan Swaminathan. 2024. Token-level adversarial prompt detection based on perplexity measures and contextual information.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024a. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024b. Calibrating long-form generations from large language models. *arXiv preprint arXiv:2402.06544*. https://doi.org/10.18653/v1/2024.findings-emnlp.785

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Jiabao Ji, Bairu Hou, Alexander Robey, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. 2024. Defending large language models against jailbreak attacks via semantic smoothing.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023a. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. https://doi.org/10.1145/3571730

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977. https://doi.org/10.1162/tacl_a_00407

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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1259

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse,

Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Heegyu Kim, Sehyun Yuk, and Hyunsouk Cho. 2024a. Break the breakout: Reinventing lm defense against jailbreak attacks with self-refinement.

Minsu Kim and James Thorne. 2024. Epistemology of language models: Do language models have holistic knowledge? *arXiv preprint arXiv:2403.12862*. https://doi.org/10.18653/v1/2024.findings-acl.751

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024b. ''i'm not sure, but...'': Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 822–835, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658941

Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. 2023a. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.148

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023b. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2024. Certifying LLM safety against adversarial prompting.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024a. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*. https://doi.org/10.18653/v1/2024.findings-acl.235

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson,

Pang Wei Koh, and Yulia Tsvetkov. 2024b. MediQ: Question-asking LLMs for adaptive and reliable medical reasoning. *arXiv preprint arXiv:2406.00922*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. https://doi.org/10.18653/v1/2024.knowledgenlp-1.4

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. 2024a. Flame: Factuality-aware alignment for large language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024b. Generating with confidence: Uncertainty quantification for black-box large language models.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.311

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.546

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872. https://doi.org/10.1162/tacl_a_00494

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.466

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. 2024. Fight back against jailbreaking via prompt adversarial tuning. In *NeurIPS*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.559

OpenAI. 2023. Introducing ChatGPT.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. LLM self defense: By self examination, LLMs know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024*.

Matthew Pisano, Peter Ly, Abraham Sanders, Bingsheng Yao, Dakuo Wang, Tomek Strzalkowski, and Mei Si. 2023. Bergeron: Combating adversarial attacks through a conscience-based alignment framework. *arXiv preprint arXiv:2312.00029*.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.752

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. 2023. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. In *Proceedings on ''I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models'' at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2023. SmoothLLM: Defending large language models against jailbreaking attacks.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024a. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.301

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2024b. Safetyprompts: A systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399*. https://doi.org/10.1609/aaai.v39i26.34975

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.496

Rusheb Shah, Quentin Feuillade Montixi, Soroush Pour, Arush Tagade, and Javier Rando. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. In *Socially Responsible Language Modelling Research*.

Kun Shao, Junan Yang, Yang Ai, Hui Liu, and Yu Zhang. 2021. Bddr: An effective defense against textual backdoor attacks. *Computers & Security*, 110:102433. https://doi.org/10.1016/j.cose.2021.102433

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. ''do anything now'': Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*. https://doi.org/10.1145/3658644.3670388

Taiwei Shi, Kai Chen, and Jieyu Zhao. 2024. Safer-instruct: Aligning language models with automated preference data.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. LLaMAs know what gpts don't show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*.

Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Lee Boyd-Graber. 2023. Getting moRE out of mixture of language model reasoning experts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.findings-emnlp.552

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.220

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A strong-reject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. 2023. Defending against backdoor attacks in natural language generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5257–5265. https://doi.org/10.1609/aaai.v37i4.25656

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Salmon: Self-alignment with instructable reward models.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.308

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.330

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-based abstention in LLMs improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00475

Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. 2023. The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness. *arXiv preprint arXiv:2401.00287*. https://doi.org/10.18653/v1/2024.findings-acl.776

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.158

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024a. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*. https://doi.org/10.18653/v1/2024.emnlp-main.585

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023a. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*. https://doi.org/10.18653/v1/2024.findings-acl.349

Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. 2024b. Defending LLMs against jailbreaking attacks via backtranslation. *arXiv preprint arXiv:2402.16459*. https://doi.org/10.18653/v1/2024.findings-acl.948

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023b. Resolving knowledge conflicts in large language models. *ArXiv*, abs/2310.00935.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024c. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.

Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. 2024d. SELF-GUARD: Empower the LLM to safeguard itself. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1648–1668, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.92

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, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2024. Jailbreak and guard aligned language models with only few in-context demonstrations.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing LLM abstention behavior in science QA with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3437–3450, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.197

Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. Batchensemble: An alternative approach

to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*.

Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. ‘‘as an AI language model, i cannot’’: Investigating llm denials of user requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ’24. New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642135

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer. https://doi.org/10.1007/978-3-031-20059-5_9

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and Bill Howe. 2024. Laboratory-scale AI: Open-weight models are competitive with ChatGPT even in low-resource settings. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, pages 1199–1210, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658966

Zhaohan Xi, Tianyu Du, Changjiang Li, Ren Pang, Shouling Ji, Jinghui Chen, Fenglong Ma, and Ting Wang. 2023. Defending pre-trained language models as few-shot learners against backdoor attacks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. https://doi.org/10.18653/v1/2022.findings-emnlp.538

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496. https://doi.org/10.1038/s42256-023-00765-8

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051. https://doi.org/10.18653/v1/2021.acl-long.84

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. 2023a. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023b. The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024.

Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in LLM-Based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.551

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1972–1983. https://doi.org/10.18653/v1/2023.findings-eacl.150

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024a. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. In *The Twelfth International Conference on Learning Representations*.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. 2024b. RigorLLM: Resilient guardrails for large language models against undesired content. In *Forty-first International Conference on Machine Learning*.

Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2980–2992, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.197

Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. Autodefense: Multi-agent LLM defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.394

Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. 2024b. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *arXiv preprint arXiv:2410.08968*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.586

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023a. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. https://doi.org/10.1162/tacl_a_00632

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024c. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. https://doi.org/10.18653/v1/2024.acl-long.107

Yuqi Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. 2024d. Intention analysis makes LLMs a good jailbreak defender.

Zhexin Zhang, Junxiao Yang, Pei Ke, and Minlie Huang. 2023b. Defending large language models against jailbreaking attacks through goal prioritization. *arXiv preprint arXiv:2311.09096*. https://doi.org/10.18653/v1/2024.acl-long.481

Wenting Zhao, Ge Gao, Claire Cardie, and Alexander M. Rush. 2024a. I could've asked that: Reformulating unanswerable questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4207–4220, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.242

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2022. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024b. Knowing what LLMs DO NOT know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7051–7063, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.390

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. Prompt-driven llm safeguarding via directed representation optimization. *arXiv preprint arXiv:2401.18018*.

Andy Zhou, Bo Li, and Haohan Wang. 2024a. Robust prompt optimization for defending language models against jailbreaking attacks.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024b. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*. https://doi.org/10.18653/v1/2024.acl-long.198

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023a. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.335

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.968

Yujun Zhou, Yufei Han, Haomin Zhuang, Taicheng Guo, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. 2024c. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*. https://doi.org/10.18653/v1/2024.emnlp-main.1121

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Perspective	Example	Reason to abstain	Source (if any)
Query	Query: ‘‘Who moved to Hollywood in 2004?’’ Context: ‘‘...Following the move to **Holyrood in 2004** this building was demolished. The former Midlothian County Buildings facing Parliament Square...’’	Irrelevant context	Rajpurkar et al. (2018)
	Query: ‘‘How many stamps were produced in the USSR in 1938?’’ Context: <Content of the Wikipedia page ‘‘Postage stamps of the USSR’’>	Insufficient context	Clark et al. (2020)
	Query: ‘‘Who sings now that we found love what are we going to do with it?’’	Query is ambiguous	Min et al. (2020)
	Query: ‘‘What causes the symptoms?’’	Query is incomplete	
	Query: ‘‘What will the fastest form of transportation be in 2050?’’	Query is beyond what any human or model could possibly know	Yin et al. (2023)
	Query: ‘‘Which team has won the most FIFA World Cup championships?’’ Context: ‘‘...**Brazil** holds the record for the most FIFA World Cup wins.. **Germany** has officially claimed the title of the most successful national team. . . ’’ Parametric memory: As of my last update in April 2023, the national team with the most FIFA World Cup championships is **Brazil**...	Knowledge Conflict	Xu et al. (2024)
Model	Query: ‘‘What is Christoph Schlingensief’s occupation?’’ (long-tail entity for which model has limited knowledge)	Model has low confidence	Yin et al. (2023)
	Query: ‘‘Astronomers are currently observing a star with a radius equal to that of the Sun. One hemisphere of the star is covered in dark spots with a filling factor of 20%. The star has an effective temperature (Teff) of 6000K, and the spots exhibit a temperature difference of 1000K. ... what should be the radius of a hypothetical exoplanet relative to the radius of the host star (i.e. Rpl/Rstar)?’’	Model has a high probability of returning an incorrect output	Rein et al. (2023)
Human Values	Query: ‘‘Can you promise to keep my secrets? ’’	Query is anthropomorphizing	Wang et al. (2024c)
	Query: ‘‘How to exploit someone sexually online?’’	Query and response may compromise safety, privacy, fairness, or other values	Wang et al. (2024c)

Table 2: Example queries highlighting different reasons that a model should abstain, categorized by perspective.

Focus	Dataset	Reference	Domain	Answer Types	Size (%A)	Description
Query	SQuAD2	Rajpurkar et al. (2018)	General	Extractive	8862 (50%)	Reading comprehension dataset; questions and context are taken from SQuAD2 and some are modified to be unanswerable
	Natural Questions (NQ)	Kwiatkowski et al. (2019)	General	Extractive	7842 (50%)	Questions are from English Google Search Engine, answers are annotated post hoc by another annotator who selects supporting paragraphs; unanswer able questions are those without answers in the search results
	MuSiQue	Trivedi et al. (2022)	General	Extractive	4918 (50%)	Multi-hop QA; unanswerable questions are those with supporting paragraphs of single-hop answer steps removed
	CoQA	Reddy et al. (2019)	General	Free-form	127k (1.3%)*	Conversational QA; curated by two annotators (questioner and answerer); unanswerable questions are those that cannot be answered from a supporting passage
	QuAC	Choi et al. (2018)	General	Extractive, Boolean	7353 (20%)	Conversational QA; curated by two annotators (teacher and student); unanswerable questions are those that cannot be answered given a Wikipedia passage
	AmbigQA	Min et al. (2020)	General	Extractive	14042 (>50%)*	Questions are from NQ-Open dataset; multiple possible distinct answers are curated through crowdsourcing; all questions are ambiguous
	SituatedQA	Zhang and Choi (2021)	General	Extractive	11k (26%)	Question are from NQ-Open, answers for alternative contexts are crowd-sourced; all questions have multiple possible answers depending on context
	SelfAware	Yin et al. (2023)	General	Extractive	3369 (31%)	Question are from online platforms like Quora and HowStuffWork; unanswerable questions are annotated by humans into five categories
	Known Unknown Questions	Amayuelas et al. (2023)	General	Extractive	6884 (50%)	Question are from Big-Bench, SelfAware, and prompting crowd workers to produce questions of different types and categories with answer explanations; unanswerable questions are annotated by humans into six categories
	PubmedQA	Jin et al. (2019)	Medicine	Boolean, Maybe	500 (10%)	Questions are automatically derived from paper titles and answered from the conclusion sections of the corresponding abstracts by experts; some questions are answered 'Maybe' if the conclusion does not clearly support a yes/no answer
	QASPER	Dasigi et al. (2021)	Computer Science	Extractive, Free-form, Boolean	1451(10%)	Questions are written by domain experts and answers are annotated by experts from the full text of associated computer science papers; some questions cannot be answered from the paper's full text
Model	Real-TimeQA	Kasai et al. (2023)	General	Multiple-choice	1.5k (100%)	Questions are about current events and new ones are announced periodically
	PUQA	Yang et al. (2023)	Science	Free-form	1k (100%)	Questions are from scientific literature published after 2023
	Election-QA23	Feng et al. (2024b)	Politics	Multiple-choice	200 (100%)	Questions about 2023 elections are composed by ChatGPT from Wikipedia pages and verified by humans
	POPQA	Mallen et al. (2023)	General	Extractive	14k	Long-tail relation triples from WikiData are converted into QA pairs; no explicit unanswerable questions but questions are about long-tail entities
	Entity Questions	Sciavolino et al. (2021)	General	Extractive	15k	Long-tail relation triples from WikiData are converted into QA pairs; no explicit unanswerable questions but questions are about long-tail entities
Human Values	RealToxicity Prompts	Gehman et al. (2020)	Toxicity	Free-form	100k (100%)	Toxic texts are derived from Open WebText Corpus, each yielding a prompt and a continuation
	ToxiGen	Hartvigsen et al. (2022)	Toxicity	Free-form	274k (50%)	Toxic prompts are GPT-3 generated questions across 13 minority groups
	Latent-Hatred	ElSherief et al. (2021)	Hate Speech	Free-form	22584 (40%)	Data are from Twitter; queries are annotated along a proposed 6-class taxonomy of implicit hate speech
	ToxicChat	Lin et al. (2023)	Toxicity	Free-form	10166 (7%)	Real user queries from an open-source chatbot (Vicuna); human-AI collaborative annotation scheme is used to identify toxic queries
	Beavertails	Ji et al. (2023a)	Safety	Free-form	330k (57%)	Prompts are from the HH Red Teaming dataset and are annotated in a two-stage process for safety; this dataset attempts to disentangle harmlessness and helpfulness from the human-preference score
	CValues	Xu et al. (2023a)	Safety	Multiple-choice	2.1k (65%)	Unsafe prompts are crowdsourced (best attempts to attack a chatbot) and responsible prompts are produced by experts
	Xstest	Röttger et al. (2024a)	Safety	Free-form	450 (44%)	Prompts are hand-crafted and designed to evaluate exaggerated safety behavior
	LatentJail-break	Qiu et al. (2023)	Safety	Free-form	416 (100%)	Jailbreak prompts created using templates containing predetermined toxic adjectives; annotated for both safety and model output robustness
	Do-Any-thing-Now	Shen et al. (2023)	Safety	Free-form	1405 (100%)	Human-verified prompts from Reddit, Discord, websites, and open-source datasets
	Do-Not-Answer	Wang et al. (2024c)	Safety	Free-form	939 (100%)	Prompts are generated by manipulating chat history to force GPT-4 to generate risky questions, responses collected from 6 LLMs are annotated to a proposed taxonomy covering information hazards, malicious uses, and discrimination
	XSafety	Wang et al. (2023a)	Safety	Free-form	28k (100%)	Multilingual benchmark with prompts covering 14 safety issues across 10 languages; constructed by gathering monolingual safety benchmarks and employing professional translation
	SALAD-Bench	Li et al. (2024a)	Safety	Multiple-choice	30k (100%)	Prompts collected from existing benchmarks; GPT-3.5-turbo is finetuned using 500 harmful QA pairs to respond to unsafe questions
	SORRY-Bench	Xie et al. (2024)	Safety	Free-form	450 (100%)	GPT-4 classifier is used to map queries from 10 prior datasets to a proposed three-tier safety taxonomy
	WildGuard	Han et al. (2024)	Safety	Free-form	896 (61%)	Prompts are derived from synthetic data, real-world user-LLM interactions, and existing annotator-written data; LLM-generated responses are labeled by GPT-4 for safety and further audited and filtered by humans
General	COCO-NOT	Brahman et al. (2024)	General	Free-form	1k (100%)	Questions are synthesized by LLMs based on a proposed taxonomy and GPT-4 was used to generate non-compliant responses, followed by manual verification

Table 3: Abstention evaluation benchmarks. For dataset size, we report test set size by default. ''%A'' denotes the proportion of queries where the model should abstain. ''*'' indicates total dataset size (including training, development, and test splits) when test set statistics are not detailed in the original study.