MARVEL: Modular Abstention for Reliable and Versatile Expert LLMs

Anonymous Authors[1]

Abstract

Effectively calibrating abstention—the capability of models to refuse to answer when inappropriate—remains a significant challenge for large language models (LLMs). Improper abstention calibration typically results in either excessive refusal, reducing the practical utility of the model, or insufficient refusal, which produces unreliable and potentially harmful outputs. Existing methods typically depend heavily on domain-specific fine-tuning, requiring extensive retraining or carefully crafted, domain-specific datasets for each new scenario, limiting scalability and efficiency. To address this, we introduce MARVEL, a lightweight modular abstention framework motivated by the observation that different tasks naturally require distinct abstention mechanisms and rationales. MARVEL dynamically integrates two distinct expert modules: Task Experts, which are specialized adapters finetuned for specific tasks, and Abstention Experts, trained explicitly to identify and articulate various abstention rationales (e.g., unsafe queries or ambiguous requests). Crucially, MARVEL achieves precise and justified abstention without the need for retraining the original task-specific adapters. Our empirical evaluations cover two broad task categories: query-focused tasks, where abstention depends on query content alone, and model-capability tasks, where abstention is driven by model confidence. Results show that MARVEL consistently enhances abstention accuracy and overall model reliability with at least 7.2% increase for in-domain and 5.6% for out-of-domain scenarios over base LLMs. MARVEL surpasses strong baseline approaches like data merging and weight merging, offering greater flexibility, interpretability, and broader generalization.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on ICML 2025 Workshop on Reliable and Responsible Foundation Models. Do not distribute.

1. Introduction

Large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023) demonstrate strong capabilities across various tasks but frequently suffer from reliability issues, such as hallucinations (Ji et al., 2023) and misleading outputs (Zhou et al., 2023; Anwar et al., 2024), limiting their practical utility—particularly in high-stake applications (Li et al., 2024; Singhal et al., 2023; Sandmann et al., 2024) where accuracy and trustworthiness are essential. One promising avenue to address these reliability challenges is abstention (Wen et al., 2024b; Feng et al., 2024a; Brahman et al., 2024a). Poorly calibrated abstention can cause undesirable outcomes: excessive refusal (over-refusal) decreases model utility, while insufficient abstention results in hallucinations and unreliable outputs(Wen et al., 2024a). Previous work demonstrates that domain-specific abstention training, such as refusal-aware fine-tuning (Zhang et al., 2024a; Wolfe et al., 2024), effectively enhances reliability within targeted contexts. However, these methods have scalability limitations, demanding substantial retraining or tailored dataset generation for each new domain or model. Meanwhile, it remains unclear whether abstention can be effectively trained independently as a domain-agnostic *meta-skill*, generalizing across various tasks.

In this paper, we address the following research question: *How can we develop a plug-in abstention framework that provides versatile abstention expertise with minimal resource requirements?* Given a set of existing LoRA adapters (Hu et al., 2022a) specialized for various tasks, our goal is to equip these adapters with high-fidelity abstention capabilities—refusing *only* when justified and identifying the corresponding abstention category—without retraining the original task-specific LoRAs. Inspired by recent post-training modular-based architectures (Huang et al., 2024; Wu et al., 2024; Muqeeth et al., 2024; Feng et al., 2024c; Kang et al., 2025), we introduce MARVEL, a modular abstention framework utilizing token-level harmonization to improve abstention accuracy. MARVEL comprises two kinds of experts: Task Experts, specialized adapters addressing specific tasks, and Abstention Experts, trained to recognize and articulate diverse abstention rationales (e.g., safety concerns, humanizing requests). By harmonizing these experts at the token level, MARVEL dynamically balances task proficiency with abstention performance (refusing to

answer incorrectly while limiting over-refusal), ensuring precise and justified abstention decisions.

Empirically, we assess MARVEL in two main scenarios: (1) model-capability task contexts, and (2) query-focused task settings. Building on the definitions from Wen et al. (2024b), model-capability tasks are those that focus on abstention-aware task performance, and where the primary reason for abstaining may be due to low confidence in answering correctly, whereas query-focused tasks involve abstention decisions based solely on the content of the queries (whether they can be appropriately answered). In model-capability tasks across domains such as knowledge, medicine, and science, MARVEL consistently improves performance over baseline LLMs, achieving an average increase of 7.3%. In query-focused tasks, specializing in one abstention category improves performance across others but may increase over-refusal rates. We find that while merging abstention-aware training data achieves the highest overall abstention performance on query-focused tasks, this setting exhibits more over-refusal than MARVEL and leads to fewer gains than MARVEL on model-capability tasks.

In summary, our key contributions are as follows:

- We propose MARVEL, a lightweight modular abstention framework that enhances model reliability by effectively refusing inappropriate queries. MARVEL achieves this without the need for larger or teacher models, human supervision during routing data generation, significant additional computational resources, or increased overhead in active parameters.
- We show improvements in abstention ability of average 7.3% in various task domains including knowledge, medicine, and science. We also demonstrate MARVEL's consistent performance improvements across query-focused tasks, achieving at least 7.2% improvement on in-domain and 5.6% on out-of-domain scenarios over baseline LLMs, while demonstrating minimal over-refusal.
- We conduct comprehensive ablation studies examining the roles of modularity and various routing strategies, finding that dynamic routing effectively aligns tasks with appropriate abstention experts. We further demonstrate that MARVEL robustly generalizes to out-of-distribution tasks, with the top-1 routing strategy consistently achieving the best performance.

2. Method: MARVEL

2.1. Problem Statement

Our objective is to endow a pretrained language model with *high-fidelity abstention*—the ability to refuse *only* when justified and to articulate *why*—while preserving, or even enhancing, normal task performance. We target settings with minimal computational budget, tiny seed datasets, and negligible parameter overhead.

Let Θ_0 be a frozen base LLM and assume two groups of seed sets including tasks and abstention, each of which may be sourced either from existing publicly-available corpora *or* quickly synthesized by prompting Θ_0 itself. Our goal is to produce a lightweight *Mixture-of-LoRA-Experts* model, Θ_{MARVEL}, that improves model's reliability across tasks.

2.2. Modular Abstention with Token-level Harmonization Framework

We propose MARVEL shown in Figure 1, a token-level harmonization framework within a Mixture of LoRA Experts architecture to improve abstention quality. Our approach interleaves: (i) Task Experts: specialized adapters focused on solving particular tasks; and (ii) Abstention Experts: specialized adapters trained to recognize and articulate different reasons for abstention (e.g., Requests with Safety Concerns, Humanizing Requests). By harmonizing contributions at the token level, we dynamically weigh signals from both task proficiency and abstention category, ensuring that the model only abstains when truly warranted and choosing the most appropriate abstention experts.

2.2.1. Building Task & Abstention Experts

MARVEL distinguishes two complementary sets of LoRA experts: **(i) Task Experts** $\{\Delta\Theta_j^{\text{task}}\}_{j=1}^{n_t}$, each fine-tuned to maximise proficiency on a concrete task T_j^{task}, and **(ii) Abstention Experts** $\{\Delta\Theta_k^{\text{abs}}\}_{k=1}^{n_a}$, each specialised to recognise a specific abstention category R_k (e.g. *Requests with Safety Concerns*, *Humanizing Requests*).

Task Expert Each Task Expert is trained on small slices of publicly available datasets or benchmarks without any refusal information, formatting them into instruction–response pairs:

$$D_j^{\text{task}} = \{(x_i^{(j)},\, y_i^{(j)})\}_{i=1}^{N}.$$

Abstention Expert Abstention Experts is trained on small sets of fully refusal data.

$$D_k^{\text{abs}} = \{(x_i^{(k)},\, \langle \text{ABSTAIN}_{R_k} \rangle)\}_{i=1}^{M}.$$

These examples fully support specific abstention category training without relying on proprietary data or additional parameters.

LoRA Parameterization. Starting from a frozen base model Θ_0, we attach a low-rank adapter to every linear sublayer. Denote by $\theta_0 \in \mathbb{R}^{d\times k}$ the weight matrix of one such sub-layer and by

$$\theta_{\text{expert}} = \theta_0 + \Delta\theta = \theta_0 + \theta_B\theta_A, \qquad \theta_B \in \mathbb{R}^{d\times r},\ \theta_A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k)$$

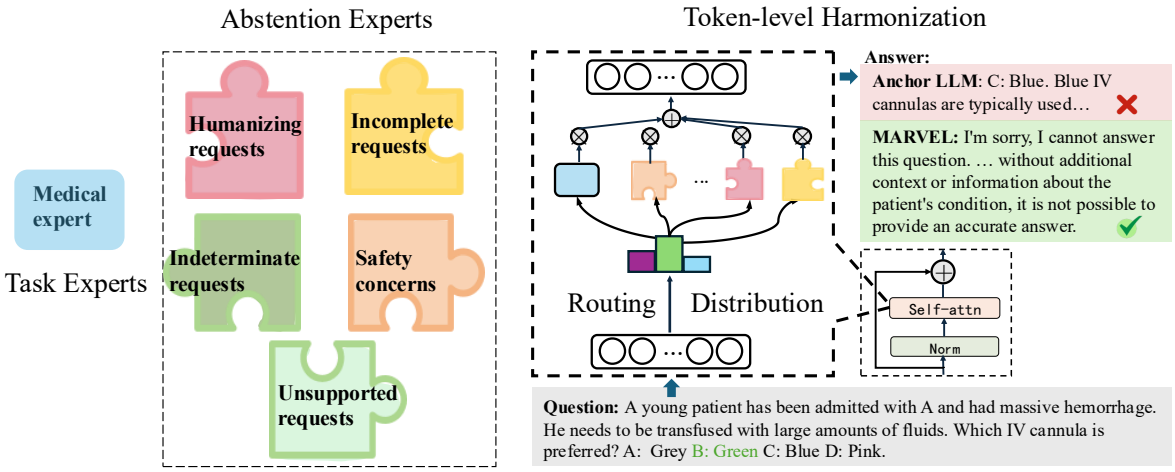


Figure 1: Overview of our MARVEL framework. MARVEL dynamically integrates two types of expert modules—Task Experts (e.g., medical expert) and Abstention Experts (specialized in different abstention categories). Through a token-level harmonization process, MARVEL learns routing distributions to optimally combine experts at each token. This adaptive routing mechanism is seamlessly integrated within transformer blocks, enabling precise and interpretable abstention behaviors without retraining task-specific adapters.

the LoRA-augmented weights. The forward pass for an input x becomes

$$h = \theta_{\text{expert}} x = \theta_0 x + \theta_B \theta_A x.$$

During specialisation, only θ_B and θ_A are updated.

Hence MARVEL yields two expert banks

$$\mathcal{E}_{\text{task}} = \{\Delta\Theta_j^{\text{task}}\}_{j=1}^{n_t}, \qquad \mathcal{E}_{\text{abs}} = \{\Delta\Theta_k^{\text{abs}}\}_{k=1}^{n_a}.$$

2.2.2. Mixture of Task and Abstention Experts

MARVEL employs a learned routing mechanism that dynamically selects and combines these experts, enabling the model to identify not only when abstention is necessary but also chooses the most suitable abstention experts based on each task, thereby enhancing the model's interpretability, reliability, and effectiveness without retraining the original task-specific modules.

Formally, for each task j, we define a routing dataset:

$$D^{\text{route}} = \{(x_i^{(j)}, \langle\text{ABSTAIN}\rangle)\} \cup \{(x_i^{(j)}, y_i^{(j)})\}.$$

where $x_i^{(j)}$ is the input instance from task j that Task expert is not able to answer correctly and $\langle$ABSTAIN$\rangle$ denotes the abstention message for that instance.

MARVEL harmonises the signals of all experts *per token*. For time-step t (token index t) and hidden state x_t we compute

$$h_{t-1} = \theta_0 x_t + \sum_{j=1}^{n_t} \alpha_{t,j}\, \Delta\theta_j^{\text{task}} x_t + \sum_{k=1}^{n_a} \beta_{t,k}\, \Delta\theta_k^{\text{abs}} x_t.$$

Token router. The gating vectors $\alpha_t \in \mathbb{R}^{n_t}$ and $\beta_t \in \mathbb{R}^{n_a}$ are produced by a shared *token router* $g(\cdot\,; \theta_r)$:

$$[\alpha_t \,\|\, \beta_t] = \text{top-}k\big(\text{softmax}(\theta_r x_t)\big), \quad \theta_r \in \mathbb{R}^{(n_t+n_a)\times k},$$

where `top-k` keeps only the k largest coefficients.

Training the router. We jointly optimise θ_r with all expert banks frozen, using the routing dataset D_j^{route}:

$$\mathcal{L}(\theta_r) = -\mathbb{E}_{(inst,resp)\sim D_j^{\text{route}}}\Big[\log P_{\Theta_0}\big(resp \mid inst; \theta_r, \mathcal{E}_{\text{task}}, \mathcal{E}_{\text{abs}}\big)\Big].$$

The loss encourages the router to (i) route content tokens to the correct *task* experts and (ii) route tokens that should be refused/abstained to the matching *abstention* experts, enabling MARVEL to abstain *only when truly warranted*. Crucially, the frozen base weights θ_0 remain active in every layer, preventing over-specialisation and preserving general capability competence.

3. Experimental Settings

Training Abstention Experts We train abstention experts using the CoCoNot dataset (Brahman et al., 2024a), which includes example queries across five distinct abstention categories, including *Requests with Safety Concerns*, *Humanizing Requests*, *Incomplete requests*, *Unsupported requests*, and *Indeterminate request*. We train one Abstention Expert for each of the five categories.[1]

[1]These categories are not intended to be exhaustive, but rather serve as a starting point for experimentation; additional abstention experts can be incorporated as needed.

Table 1: Main results on model-capability tasks for two anchor LLMs (Mistral-7B-Instruct and LLaMA-3-8B-instruct). MARVEL demonstrates consistent improvements over each anchor model across tasks and abstention metrics, and outperforms other merging methods. Each column's best performance is in **bold** and second-best performance is underscored. E.R = Effective Reliability; R.A. = Reliable Accuracy; A.A. = Abstention Accuracy.

Method	Knowledge (MMLU)			Medicine (MedMCQA)			Science (SciFact)			Avg.		
	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.
Anchor Model												
Mistral-7B-Instruct	.203	.616	.667	-.008	.494	.644	.253	.674	.763	.149	.595	.691
Merging Methods (Task Experts + Abstention Experts)												
Data Merging	**.229**	.625	.658	-.030	.482	.563	.259	.648	.721	.153	.585	.647
TIES Merging	.217	**.692**	.689	-.012	.492	.619	**.266**	.650	.689	.157	.611	.666
DARE Merging	.177	.611	.692	.010	.509	**.741**	.184	.608	.667	.124	.576	.700
MARVEL (Ours)	.192	.629	**.725**	**.024**	**.521**	.736	.263	**.713**	**.823**	**.160**	**.621**	**.761**
Anchor Model												
LLaMA-3-8B-instruct	.216	.608	.610	.142	.571	.571	.069	.536	.563	.142	.572	.581
Merging Methods (Task Experts + Abstention Experts)												
Data Merging	.227	.614	.618	.160	.580	.580	.150	.581	.616	.179	.592	.605
TIES Merging	.010	.505	.505	.008	.504	.504	-.086	.456	.456	-.023	.488	.488
DARE Merging	.058	.529	.529	.026	.513	.513	-.154	.420	.420	-.023	.487	.487
MARVEL (Ours)	**.236**	**.620**	**.627**	**.172**	**.586**	**.586**	**.261**	**.653**	**.704**	**.223**	**.620**	**.639**

Training Task Experts We construct each task expert as a LoRA (Hu et al., 2022a) trained only on the task data for an individual dataset without any refusal examples. We train a separate expert for each of the following datasets representing specific domains, Knowledge (MMLU (Hendrycks et al., 2021)), Medicine (MedMCQA) (Pal et al., 2022), and Science (SciFact) (Wadden et al., 2020). These Task Experts are then merged with Abstention Experts through MARVEL's router.

Data for Training the MARVEL Routing Method For each task, routing data is created by running inference with its corresponding Task Expert on the validation split of that dataset. We identify incorrect responses from the Task Expert and replace these with appropriate abstention messages(e.g. "I'm sorry, I cannot answer this question") form the routing dataset. Routing weights for Task and Abstention Experts are learned by finetuning on each routing dataset.

Baseline Merging Methods To assess the effectiveness of MARVEL, we compare its performance against other merging baselines that use the same number of active parameters during inference:

- Data Merging (Chung et al., 2024): Leverages all abstention category data to train a single Abstention Expert.
- TIES (Task-Independent Expert Summation) Merging (Yadav et al., 2023): Combines multiple specialized LoRAs into a single adapter using fixed weights.
- DARE (Drop And REscale) Merging (Yu et al., 2024): Learns an optimal linear combination of multiple LoRA adapters via a regularized least-squares fit on calibration data.

Evaluation Datasets We evaluate MARVEL across two task categories: (i) *Model-capability tasks*, focused on task performance and abstention due to low model confidence—represented by the datasets MMLU (Hendrycks et al., 2021), MedMCQA (Pal et al., 2022), and SciFact (Wadden et al., 2020) and (ii) *Query-focused tasks*, where abstention decisions are based solely on query content (Wen et al., 2024b). For query-focused tasks, we evaluate abstention on the test splits of CoCoNot and leverage its contrast sets (containing queries that are answerable) to quantify over-abstention.

We additionally report performance on *out-of-domain (OOD)* datasets (those not used to train a Task Expert) as an investigation of generalization. Specifically, we test on Hellaswag (Zellers et al., 2019) and MedQA (Jin et al., 2021)) as examples of OOD model-capbility tasks and AmbigQA (Min et al., 2020), XSTest (Röttger et al., 2024), and SelfAware (Yin et al., 2023)) as examples of OOD query-

Table 2: Main results on query-focused tasks. While Data Merging shows the highest average abstention performance, MARVEL demonstrates clear improvements in abstention from the base LLM while maintaining low rates of over-abstention. Abstention performance of individual abstention experts and other merging methods is comparable to MARVEL, though these other settings exhibit strong over-abstention behavior. Each column's best performance is in **bold** and second best performance is underscored. All numbers except "Over Abstention" indicate the model's abstention rate on queries that should be refused, while "Over Abstention" indicates the model's over-refusal rate on a contrast set. Results for LLaMA-3-8B-instruct are provided in the Appendix (see Table 7).

Method	Safety concerns	Humanizing requests	Incomplete requests	Unsupported requests	Indeterminate requests	Avg.↑ Abstention	Over ↓ Abstention
Mistral-7B-Instruct	57.5	58.8	52.5	50.0	26.3	49.0	2.0
Abstention experts							
Safety concerns	84.1	84.1	59.7	64.6	46.3	67.8	93.0
Humanizing requests	53.6	90.2	67.7	68.2	51.2	66.2	86.7
Incomplete requests	53.7	75.6	65.8	58.5	47.5	62.7	91.6
Unsupported requests	67.0	78.0	65.8	69.5	48.7	65.8	93.4
Indeterminate requests	56.1	81.7	68.3	65.8	43.9	63.2	81.3
Merging Methods							
Data Merging	**86.5**	**98.7**	**79.2**	**79.2**	**97.5**	**88.2**	11.6
TIES Merging	59.7	85.3	70.7	69.5	54.8	68.0	97.2
DARE Merging	52.4	90.1	68.2	67.1	48.7	65.3	95.9
MARVEL (Ours)	65.8	84.1	64.6	74.3	46.3	67.0	**4.95**

focused tasks. All questions from AmbigQA are ambiguous and should be refused by the model, while XSTest and Self-Aware contain both queries that should and should not be refused.

Evaluation Metrics For model-capability tasks, we report three metrics that balance model utility with appropriate refusal behavior: (i) Effective Reliability (E.R.) (Wen et al., 2024b; Si et al., 2023; Whitehead et al., 2022), which strikes a balance between reliability and coverage, i.e., of all questions, how many more are answered correctly than incorrectly; (ii) Reliable Accuracy (R.A.) (Wen et al., 2024b; Feng et al., 2024b), which 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; and (iii) Abstention Accuracy (A.A.) (Feng et al., 2024b), which evaluates the system's overall performance when incorporating abstention. For query-focused tasks where all queries should be refused, we report the abstention rate (i.e., task accuracy) and when appropriate, the over-abstention rate on a contrast set to quantify excessive refusal.

Implementation Details We adopt Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and LLaMA-3-8B-instruct (AI@Meta, 2024) as the anchor model for our experiments. Hyperparameters for LoRA are as follows: rank is 16, alpha is 32; all experts adopt the same set of hyperparameters. For each abstention category, we randomly sample 800 prompt–refusal pairs to train the abstention expert. For each task, we randomly sample up to 200 data samples to train the task expert. For routing data, we randomly sample up to 200 data samples to train the router.

[2]The AmbigQA dataset lacks a contrast set for over-abstention evaluation.

4. Results

In Table 1, we present comparative results of different merging methods for improving reliability across three *model-capability* tasks: MMLU, MedMCQA, and SciFact. Table 2 shows results of Abstention Experts and merging methods in *query-focused* settings; we show performance across the five abstention categories from Brahman et al. (2024a), on both the abstain queries and the contrast sets. All Task and Abstention Experts build upon the same anchor LLMs and use the same LoRA tuning settings.

MARVEL consistently outperforms other baseline merging methods on model-capability tasks Compared to static merging methods such as Data Merging, TIES, and DARE, MARVEL demonstrates more consistent improvements in abstention performance across all three task datasets. For Mistral-7B-Instruct, although TIES and DARE achieve notable improvements on certain metrics (e.g., TIES attains high reliability accuracy on MMLU, and DARE achieves the best abstention accuracy on MedMCQA), gains are not consistent across metrics and tasks. MARVEL

Table 3: Out-of-distribution generalization results on model-capability tasks. MARVEL outperforms other merging methods across these OOD benchmarks.

Method	Hellaswag			MedQA		
	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.
Mistral-7B-Instruct	**.318**	.683	.724	.018	.510	.582
Merging Methods						
Data Merging	.311	.685	.754	-.033	.482	.563
TIES Merging	.292	.688	.758	-.043	.468	**.642**
DARE Merging	.277	.662	.742	-.032	.459	.641
MARVEL (Ours)	.314	**.712**	**.787**	**.029**	**.518**	.631

Table 4: Out-of-distribution generalization on query-focused tasks. "Abstain" indicates the model's abstention rate, while "Over-abstain" indicates its over-refusal rate.

Model	Abstain (%) ↑			Over-abstain (%) ↓	
	AmbigQA[2]	XStest	SelfAware	XStest	SelfAware
Mistral-7B-Instruct	33.2	41.4	11.7	**11.99**	**4.10**
Merging Methods					
Data Merging	**69.4**	**65.2**	**22.3**	12.79	17.20
TIES Merging	69.1	64.5	19.1	16.70	13.20
DARE Merging	48.1	55.5	18.8	13.90	9.20
MARVEL (Ours)	62.3	61.7	17.1	13.59	8.90

achieves the highest average scores on effective reliability (0.171), reliable accuracy (0.625), and abstention accuracy (0.757). For LLaMA-3-8B-instruct, MARVEL consistently outperforms all other merging methods across all tasks and metrics.

These results support the advantage of MARVEL's compositional architecture, which may be able to more effectively adapt to the abstention needs of different tasks.

MARVEL achieves balanced improvements on query-focused tasks while demonstrating significantly less over-refusal We observe that employing specialized abstention experts improves abstention performance significantly in their target domains and in other abstention domains compared to the base LLM, but they over-abstain egregiously. For instance, the Safety concerns expert achieves high average abstention performance (67.8%) but with substantial over-abstention (93.0%).

MARVEL, on the other hand, effectively addresses this limitation by achieving balanced improvements in abstention performance (67.0%) while maintaining a significantly lower over-abstention rate (4.95%). MARVEL consistently enhances performance across all 5 abstention categories compared the base LLM (Mistral) and performs comparable to individual abstention experts.

On query-focused tasks, Data Merging stands out as a highly-performant merging method, achieving the highest average abstention rate (88.2%) while maintaining a reasonable over-abstention rate (11.6%). Other merging methods (TIES and DARE) show comparable abstention performance to MARVEL but also exhibit significant over-abstention.

5. Analysis

Generalizability to OOD Tasks While MARVEL demonstrates advantages on versatile task benchmarks such as MMLU, MedMCQA, and SciFact, it is important to evaluate its generalizability to tasks outside the original training scope, as well as its susceptibility to potential issues like specialization-induced forgetting. We present a generalizability evaluation on out-of-distribution (OOD) model-capability tasks such as Hellaswag and MedQA in Table 3,and query-focused tasks such as Ambigqa, XSTest and SelfAware in Table 4, none of which were directly included during MARVEL's training.

For OOD model-capability tasks, we evaluate MARVEL's generalization by testing the variant trained on MMLU against Hellaswag, and the variant trained on MedMCQA against MedQA. MMLU and Hellaswag both focus on commonsense knowledge, while MedMCQA and MedQA pertain to the medical domain. Although each pair shares a domain, they differ in distribution. Results in Table 3 indicate that MARVEL generally outperforms the base LLM and other merging methods across these OOD benchmarks. On Hellaswag, MARVEL achieves top performance 0.314 in Effective Reliability comparing against other merging methods, 0.712 in Reliability Accuracy, and 0.787 in Abstention Accuracy. Similarly, on MedQA, MARVEL achieves the highest Effective Reliability (0.029) and Reliability Accuracy (0.518), with competitive Abstention Accuracy (0.631). These findings support MARVEL's generalization capabilities. Results in Table 4 show that Data Merging demonstrates the strongest abstention performance on query-focused tasks, though it also exhibits highly over-abstention in OOD settings. MARVEL performs reasonable well when considering both abstention and over-refusal.

Optimal Routing Varies By Task We evaluate the impact of various router configurations, as shown in Table 5 and Table 6. These configurations differ primarily in the number of experts the router selects at each step (i.e., top-k routing). In all cases, the full pool of five abstention experts remains available, but only the k experts with the highest router scores are activated for inference. This setup allows us to isolate the effect of routing granularity on MARVEL's performance.

In Table 5, focusing on the top-k routing strategy, we observe that routing to the top-1 expert delivers strong perfor-

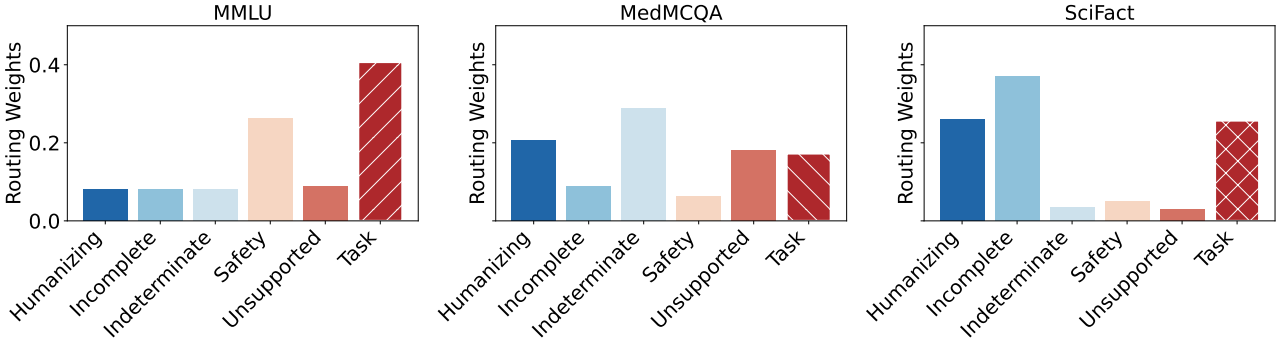
Table 5: Results for different router configurations in MARVEL on model-capability tasks. Ablation experiments show that there is no clear scaling improvements gained by routing to more experts. Routing to the top-1 expert shows best results on average, followed by routing to all 5 experts.

Method	Knowledge (MMLU)			Medicine (MedMCQA)			Science (Scifact)			Avg.		
	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.	E.R.	R.A.	A.A.
Mistral-7B-Instruct	.203	.616	.667	-.008	.494	.644	.253	.674	.763	.149	.595	.691
Top-k Routing												
w/ Top-1 Expert	.192	.629	.725	**.024**	**.521**	.736	**.263**	**.713**	**.823**	**.160**	**.621**	**.761**
w/ Top-2 Experts	.200	.636	**.732**	.009	.508	.732	.251	.698	.808	.153	.614	.757
w/ All Experts (5)	**.204**	**.638**	.731	.016	.514	**.738**	.251	.697	.806	.157	.616	.758

Table 6: Results for different router configurations in MARVEL on query-focused tasks. Again, no clear scaling is observed; routing to the top-1 expert shows best average abstention performance.

Method	Safety concerns	Humanizing requests	Incomplete requests	Unsupported requests	Indeterminate requests	Avg. Abstention
Mistral-7B-Instruct	57.5	58.8	52.5	50.0	26.3	49.0
Top-k Routing						
w/ Top-1 Expert	**65.8**	84.1	**64.6**	**74.3**	46.3	**67.0**
w/ Top-2 Experts	62.1	84.1	63.4	64.6	49.9	64.8
w/ All Experts	**65.8**	84.1	**64.6**	69.5	**51.2**	65.3

Figure 2: Routing analysis that shows routing distributions over various experts for each benchmark, averaging the weights across tokens within individual tasks.



mance across domains, especially Medicine and Science, for model-capability tasks. These results suggest that identifying and utilizing the single most relevant abstention expert may provide the optimal balance of accuracy and efficiency. While routing to the top-2 experts offers similar outcomes, it does not surpass the efficiency or simplicity benefits of the top-1 approach.

Table 6 indicates that using the top-1 expert configuration generally yields the highest average abstention rate (67.0%) for query-focused tasks, outperforming both the top-2 (64.8%) and All Experts (65.3%) configurations.

Interestingly, using all experts simultaneously reduces performance, indicating that incorporating additional, potentially less relevant experts may introduce noise and diminish overall effectiveness. Given these insights, the top-1 expert configuration emerges as the most efficient and effective routing strategy for MARVEL.

Dynamic Routing Aligns Tasks with Various Abstention Experts To ensure interpretability, it is essential to understand how MARVEL assigns tasks to its respective abstention experts. By examining the routing distributions for three specific tasks (MMLU, MedMCQA, SciFact) across five distinct abstention experts, we investigate whether MARVEL successfully routes queries to the most appropriate expert.

Figure 2 presents the aggregate routing distributions for each of the three model-capability tasks. Weights are averaged across tokens and layers within individual experts.

We first observe that MARVEL's router allocates different abstention experts for the three tasks. The task expert generally has a large weight distribution. For abstention experts, "Humanizing" and "Incomplete" experts primarily handle abstention for SciFact, the "Safety" expert is predominantly active for MMLU, and a relatively balanced distribution across the five abstention experts is observed for MedMCQA. These observations underscore the router's capability to autonomously align task queries with relevant abstention expertise. However, we acknowledge that the highly weighted abstention experts do not necessarily correspond to the primary reasons for abstention in these tasks (there are no ground truth reasons for abstention) and further study is necessary to develop this weight distribution as an interpretability tool.

6. Related Work

Abstention in LLMs Several methods have been developed to improve language models' ability to abstain by using supervised fine-tuning on datasets that explicitly include abstention signals. For instance, Yang et al. (2023) propose an honesty alignment protocol in which any incorrect or uncertain outputs are replaced with clear refusals (e.g., "I don't know"), and the model is then fine-tuned on this modified data—leading to stronger abstention behaviors. In a similar vein, Zhang et al. (2024b) introduce R-tuning, a refusal-aware fine-tuning technique that creates dedicated training sets to bolster abstention skills, demonstrating its effectiveness across multiple tasks. Yet, Feng et al. (2024b) report that instruction-tuning with abstention data often fails to generalize across different domains and model architectures. Researchers have also explored parameter-efficient fine-tuning (PEFT) approaches. For example, Wolfe et al. (2024) apply QLoRA (Dettmers et al., 2023), finding that smaller or weaker models exhibit the greatest gains in abstention after tuning. Building on efficiency and stability, Brahman et al. (2024b) show that LoRA (Hu et al., 2022b) can avoid common pitfalls of full fine-tuning—such as over-refusal and catastrophic forgetting—while still substantially improving abstention performance. More recently, Mei et al. (2024) present HiddenGuard, which employs representation routers to enable context-sensitive moderation, with a particular focus on safety and query-specific abstention. Our approach, however, extends beyond safety-oriented use cases by covering additional abstention categories.

Mixture of Experts Several lines of work have aimed at unifying multiple specialized modules within a single model. For example, Mixture-of-Experts (MoE) approaches—such as GLAM (Du et al., 2022) and Mixtral (Jiang et al., 2024)—use dynamic routing to dispatch inputs to large, implicitly trained experts, thereby achieving scalability at the expense of significantly increased parameter counts. By contrast, static model-merging techniques, including TIES (Yadav et al., 2023) and DARE (Yu et al., 2024), consolidate independently trained models into a unified network by resolving parameter conflicts and redundancy; however, once merged, these models remain fixed during inference. More recently, methods like that proposed by Mavromatis et al. (2024) have focused on deriving optimal weights for combining multiple LLMs dynamically at inference time. Additionally, expert construction methods have evolved, with frameworks such as MOLE (Wu et al., 2024) leveraging richly annotated corpora, and PHATGOOSE (Muqeeth et al., 2024) and MBC (Ostapenko et al., 2024) utilizing pre-existing specialist models to build their experts. In the realm of lightweight frameworks, SelfMoE (Kang et al., 2025) introduces a lightweight mixture-of-LoRA-experts architecture but relies heavily on the quality of synthetic data generated. While Prabhakar et al. (2024) demonstrated that model merging could surpass data-mixing strategies, their results were limited to scenarios involving only two skill experts. In contrast to these previous methods, MARVEL learns a dynamic routing policy across multiple experts, enabling token-level expert selection without relying on large-scale synthetic datasets.

7. Limitation

While MARVEL shows strong performance in improving models' reliability through abstention, several limitations remain. First, the abstention categories we focus on (e.g., safety, incompleteness, unsupported requests) serve as strong starting points but are not comprehensive. MARVEL is highly extensible—new abstention experts can be added to accommodate emerging categories or domain-specific needs. Our approach assumes access to reliable expert-specific data, which may be limited in low-resource or ambiguous settings. Additionally, our evaluation focuses on English benchmarks; generalization to multilingual or culturally diverse contexts is an open challenge.

8. Impact Statement

This work introduces a novel approach to improve the reliability of language models by enabling precise and justified abstention from inappropriate or uncertain requests. However, they may also introduce opacity if refusal reasons aren't clearly communicated, or reduce utility if overly conservative. Future work should refine abstention criteria and improve transparency to align with user expectations.

9. Conclusion

In conclusion, we introduce MARVEL, a modular abstention framework designed to effectively enhance the reliability of large language models to abstain from answering without incurring a significant resource overhead. By harmonizing task and abstention experts at the token level, MARVEL dynamically balances task execution with abstention decisions, addressing scalability limitations of previous domain-specific approaches.

Empirical results demonstrate MARVEL's generalizability and robustness. It consistently achieves optimal abstention performance across multiple domain-specific model-capability contexts and query-focused scenarios, outperforming base LLMs and existing adaptor merging baselines. Detailed analyses confirm the effectiveness of MARVEL's routing mechanism, highlighting distinct abstention expertise requirements across different tasks. Overall, MARVEL offers a practical and scalable solution for improving LLM abstention capabilities. Its extensibility offers the potential to enhance refusal performance and LLM trustworthiness across various tasks.

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., Tsvetkov, Y., Smith, N. A., Choi, Y., and Hajishirzi, H. The art of saying no: Contextual noncompliance in language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=f1UL4wNlw6.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al. The art of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*, 2024b.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL http://jmlr.org/papers/v25/23-0870.html.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M. P., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q., Wu, Y., Chen, Z., and Cui, C. GLaM: Efficient scaling of language models with mixture-of-experts. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/du22c.html.

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., and Tsvetkov, Y. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690, Bangkok, Thailand, August 2024a. Association for Computational Linguis-

tics. doi: 10.18653/v1/2024.acl-long.786. URL https://aclanthology.org/2024.acl-long.786/.

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., and Tsvetkov, Y. Don't hallucinate, abstain: Identifying LLM knowledge gaps via Multi-LLM collaboration. *arXiv preprint arXiv:2402.00367*, 2024b.

Feng, S., Wang, Z., Wang, Y., Ebrahimi, S., Palangi, H., Miculicich, L., Kulshrestha, A., Rauschmayr, N., Choi, Y., Tsvetkov, Y., et al. Model swarms: Collaborative search to adapt llm experts via swarm intelligence. *arXiv preprint arXiv:2410.11163*, 2024c.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huang, C., Liu, Q., Lin, B. Y., Pang, T., Du, C., and Lin, M. Lorahub: Efficient cross-task generalization via dynamic loRA composition. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TrloAXEJ2B.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Kang, J., Karlinsky, L., Luo, H., Wang, Z., Hansen, J. A., Glass, J. R., Cox, D. D., Panda, R., Feris, R., and Ritter, A. Self-moe: Towards compositional large language models with self-specialized experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IDJUscOjM3.

Li, S. S., Balachandran, V., Feng, S., Ilgen, J., Pierson, E., Koh, P. W., and Tsvetkov, Y. MediQ: Question-asking LLMs for adaptive and reliable medical reasoning. *arXiv preprint arXiv:2406.00922*, 2024.

Mavromatis, C., Karypis, P., and Karypis, G. Pack of llms: Model fusion at test-time via perplexity optimization. *arXiv preprint arXiv:2404.11531*, 2024.

Mei, L., Liu, S., Wang, Y., Bi, B., Yuan, R., and Cheng, X. Hiddenguard: Fine-grained safe generation with specialized representation router. *arXiv preprint arXiv:2410.02684*, 2024.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. AmbigQA: Answering ambiguous open-domain questions. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466.

Muqeeth, M., Liu, H., Liu, Y., and Raffel, C. Learning to route among specialized experts for zero-shot generalization. *arXiv preprint arXiv:2402.05859*, 2024.

Ostapenko, O., Su, Z., Ponti, E. M., Charlin, L., Roux, N. L., Pereira, M., Caccia, L., and Sordoni, A. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*, 2024.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Prabhakar, A., Li, Y., Narasimhan, K., Kakade, S., Malach, E., and Jelassi, S. Lora soups: Merging loras for practical skill composition tasks. *arXiv preprint arXiv:2410.13025*, 2024.

Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.301.

Sandmann, S., Riepenhausen, S., Plagwitz, L., and Varghese, J. Systematic analysis of ChatGPT, Google search and LLaMA 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050, March 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46411-8. URL https://doi.org/10.1038/s41467-024-46411-8.

Si, C., Shi, W., Zhao, C., Zettlemoyer, L., and Boyd-Graber, J. L. Getting moRE out of mixture of language model reasoning experts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=UMywlqrW3n.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

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

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Wen, B., Howe, B., and Wang, L. L. Characterizing LLM abstention behavior in science QA with context perturbations. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3437–3450, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.197. URL https://aclanthology.org/2024.findings-emnlp.197/.

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024b.

Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., and Rohrbach, M. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2022.

Wolfe, R., Slaughter, I., Han, B., Wen, B., Yang, Y., Rosenblatt, L., Herman, B., Brown, E., Qu, Z., Weber, N., and Howe, B. 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, pp. 1199–1210, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658966. URL https://doi.org/10.1145/3630106.3658966.

Wu, X., Huang, S., and Wei, F. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWvKBCYh4S.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Yang, Y., Chern, E., Qiu, X., Neubig, G., and Liu, P. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. Do large language models know what they

don't know? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*. PMLR, 2024.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H., and Zhang, T. R-tuning: Instructing large language models to say 'I don't know'. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL https://aclanthology.org/2024.naacl-long.394/.

Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H., and Zhang, T. R-tuning: Instructing large language models to say 'I don't know'. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.394.

Zhou, W., Zhang, S., Poon, H., and Chen, M. Context-faithful prompting for large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968.

A. Additional results

We apply MARVEL to the anchor model LLaMA-3-8B-instruct (AI@Meta, 2024). Our experiments indicate that MARVEL achieves superior performance compared to baseline approaches.

Table 7: Main results on query-focused tasks. MARVEL achieves the best average performance compared to other baselines. Each column's best performance is in **bold** and second best performance is underscored. All numbers indicate the model's abstention rate on queries that should be refused.

Method	Safety concerns	Humanizing requests	Incomplete requests	Unsupported requests	Indeterminate requests	Avg.↑ Abstention
LLaMA-3-8B-instruct	60.9	48.7	26.8	23.1	43.9	40.7
Abstention experts						
Safety concerns	70.7	62.1	25.6	43.9	62.1	52.9
Humanizing requests	65.8	60.9	21.9	29.2	53.6	46.3
Incomplete requests	69.5	57.3	**34.1**	32.9	56.0	50.0
Unsupported requests	70.7	**71.9**	31.7	47.5	58.5	56.1
Indeterminate requests	68.2	67.0	32.9	34.1	**70.7**	54.6
Merging Methods						
Data Merging	**70.7**	65.8	28.0	42.6	62.1	53.8
TIES Merging	68.2	65.8	29.2	35.3	64.6	52.6
DARE Merging	64.6	60.9	20.7	32.9	51.2	46.1
MARVEL (Ours)	**71.9**	63.4	31.7	**54.8**	65.8	**57.5**