

A Details about Datasets

The details of the 15 datasets are further listed in Tabel 2. We applied the original dataset directly from the Huggingface dataset repositories without any further processing. A thorough examination of each dataset’s attributes, size, and notable characteristics is provided below.

Table 2: Details of the 15 datasets used in our benchmark. FF: Free-form question answering (including numerical answers for math tasks); MCQ: Multiple-choice question answering; TF: True/False question answering.

Dataset	Type	Domain	# Train	# Test	Description
GSM8K	FF	Mathematical Reasoning	7473	1319	Grade school math word problems
AQuA	MCQ	Mathematical Reasoning	97467	254	Algebraic word problems
MultiArith	FF	Mathematical Reasoning	420	180	Algebraic word problems
SVAMP	FF	Mathematical Reasoning	700	300	Algebraic word problems
MATH-500	FF	Mathematical Reasoning	—	500	Algebraic word problems
BoolQ	TF	Commonsense Reasoning	9427	3270	Commonsense and factual reasoning questions
CommonsenseQA	MCQ	Commonsense Reasoning	9741	1221	Questions assessing various types of commonsense knowledge
HellaSwag	MCQ	Commonsense Reasoning	39905	10042	Sentence completion based on narrative understanding
OpenBookQA	MCQ	Commonsense Reasoning	4957	500	Open-book science and commonsense questions
PIQA	MCQ	Commonsense Reasoning	16113	1838	Physical commonsense reasoning questions
Social IQa	MCQ	Commonsense Reasoning	33410	1954	Social commonsense intelligence questions
TruthfulQA	FF	Commonsense Reasoning	653	164	Assessing models’ ability to prevent false information
WinoGrande	MCQ	Commonsense Reasoning	2558	1267	Pronoun ambiguity resolution with commonsense reasoning
CoQA	FF	Conversational & Contextual Understanding	7199	500	Conversational questions on text passages from diverse domains
MMLU	MCQ	Problem Solving	99842	14042	Problem solving across various subjects

B Related Work

B.1 Small Language Models

Small Language Models (SLMs) are designed for deployment on resource-constrained devices like desktops, smartphones, and wearables. Specifically, we consider the Transformer-based SLMs in this work due to their state-of-the-art performance, like Phi-3-mini [1], TinyLlama [81], MobileLLM [51], and Qwen-1.5B [5], LiteLLaMa-460M, OPT-125M [82], BLOOMZ (560M, 1.1B, 1.7B, 3B) [41], SmolLM (135M, 360M, 1.7B) [3], OLMo (1B) [28], OLMoE (1B) [58], MobileLlama (0.5B, 1B) [68], MobileLLaMA (1.4B, 2.7B) [10], OpenLLaMA (3B) [25]. These models are designed with lightweight architectures to operate effectively within the computational and storage limitations of mobile devices and edge hardware.

Recurrent Neural Networks (RNNs), like RWKV (1B, 3B, 7B) [61], Mamba (1.4B, 6.9B) [14], and RecurrentGemma-2B [27], can provide promising solutions for on-device inference in resource-constrained environments. These models leverage the recurrent nature of RNNs to process sequential data efficiently without requiring a KV cache, which is suitable for resource-constrained on edge devices. Specifically, RWKV introduces a hybrid RNN-Transformer backbone to capture long-term dependencies while maintaining computational efficiency. Similarly, Mamba and RecurrentGemma design recurrent layers for low-power consumption and high throughput inference, which can significantly reduce memory and computational requirements, fostering low-latency applications directly on devices.

C Additional Experimental Results from Benchmarking to Generalization

In this section, we present additional experimental results on (1) evaluating the impact of uncertainty-correctness alignment on small language model (SLM) routing and (2) investigating the generalization capability of proxy routing data on novel datasets. **Since our studies yield over 5,600 results, we here present a representative subset in the following section. The full set of results is provided in the supplementary materials.**

For the first experiment (Section C.1 and Section C.2), we provide the complete set of results, including the AUC measurements for uncertainty-correctness alignment and the performance of uncertainty-based routing. For the second experiment (Section C.3), we present a comprehensive experimental results of proxy routing prediction under partially in-domain setting. Each dataset referenced in the experiments is treated as a novel dataset for evaluation.

C.1 Evaluation on Uncertainty-correctness Alignment

Results of Alignment between uncertainty and correctness.

All the experiments shown on this page are conducted under AQUA, BoolQ, and CoQA datasets with all 8 UQ methods.

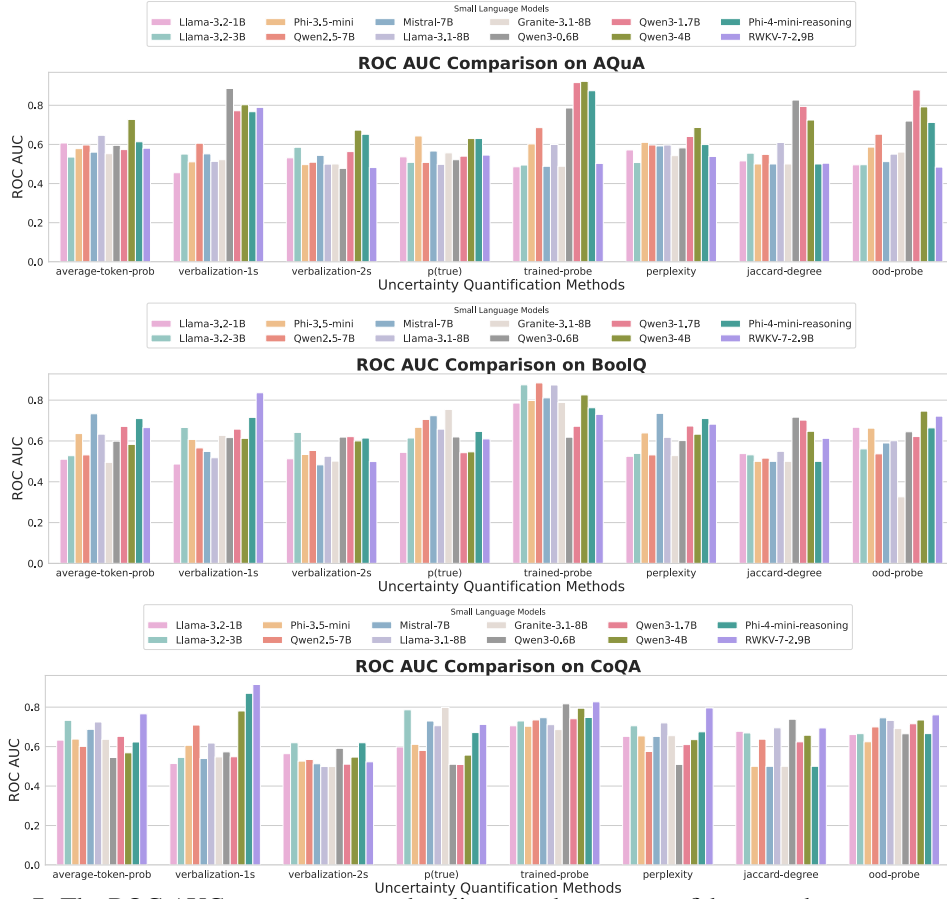


Figure 7: The ROC AUC scores measure the alignment between confidence and correctness across different SLMs and uncertainty quantification methods on AQUA, BoolQ, and CoQA. A higher ROC AUC indicates a stronger alignment.

647 **Results of Alignment between uncertainty and correctness.**

648 All the experiments shown on this page are conducted under GSM8K, HellaSwag, MMLU, and
 649 MultiArith datasets with all 8 UQ methods.

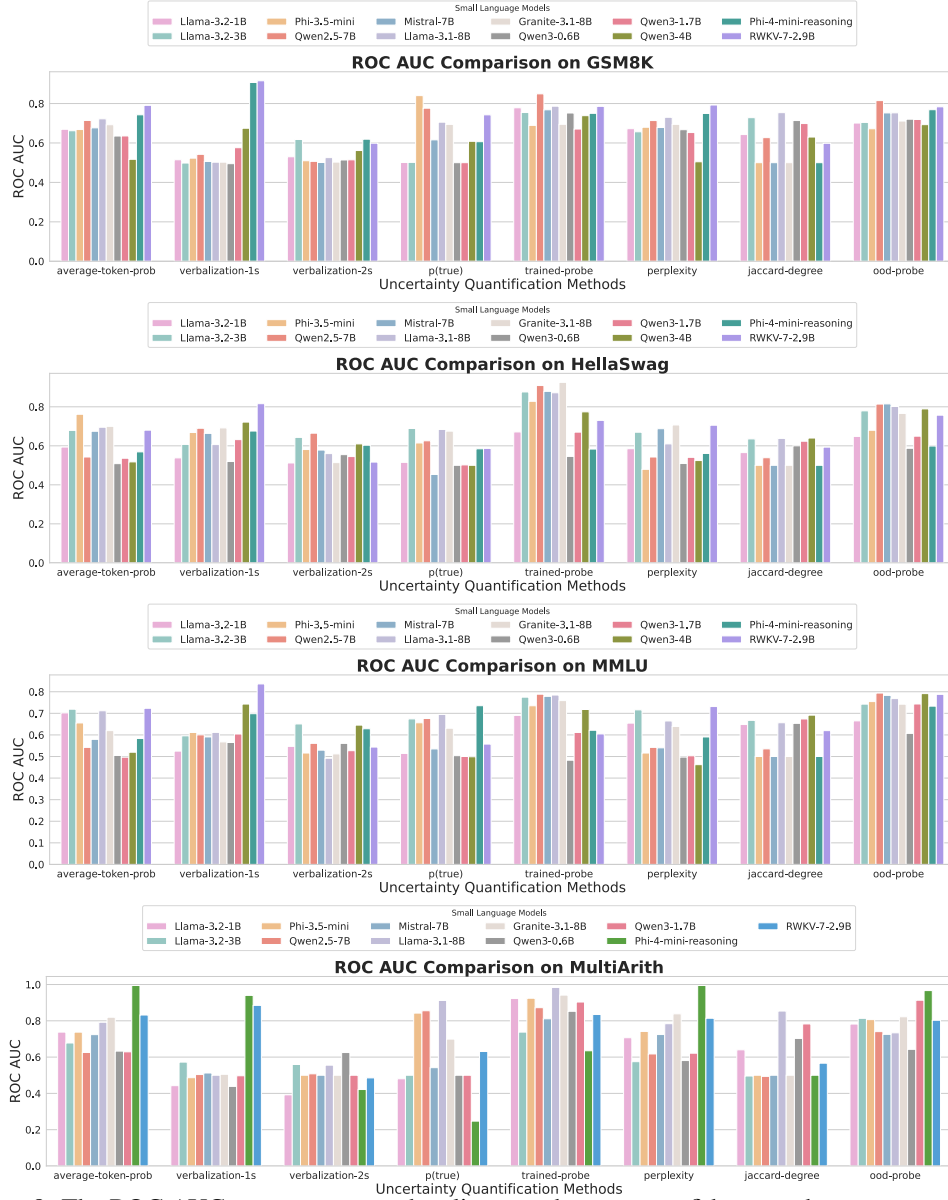


Figure 8: The ROC AUC scores measure the alignment between confidence and correctness across different SLMs and uncertainty quantification methods on GSM8K, HellaSwag, MMLU, and Multi-Arith. A higher ROC AUC indicates a stronger alignment.

650 **Results of Alignment between uncertainty and correctness.**

651 All the experiments shown on this page are conducted under OpenBookQA, PIQA, SocialIQA, and
 652 SVAMP datasets with all 8 UQ methods.

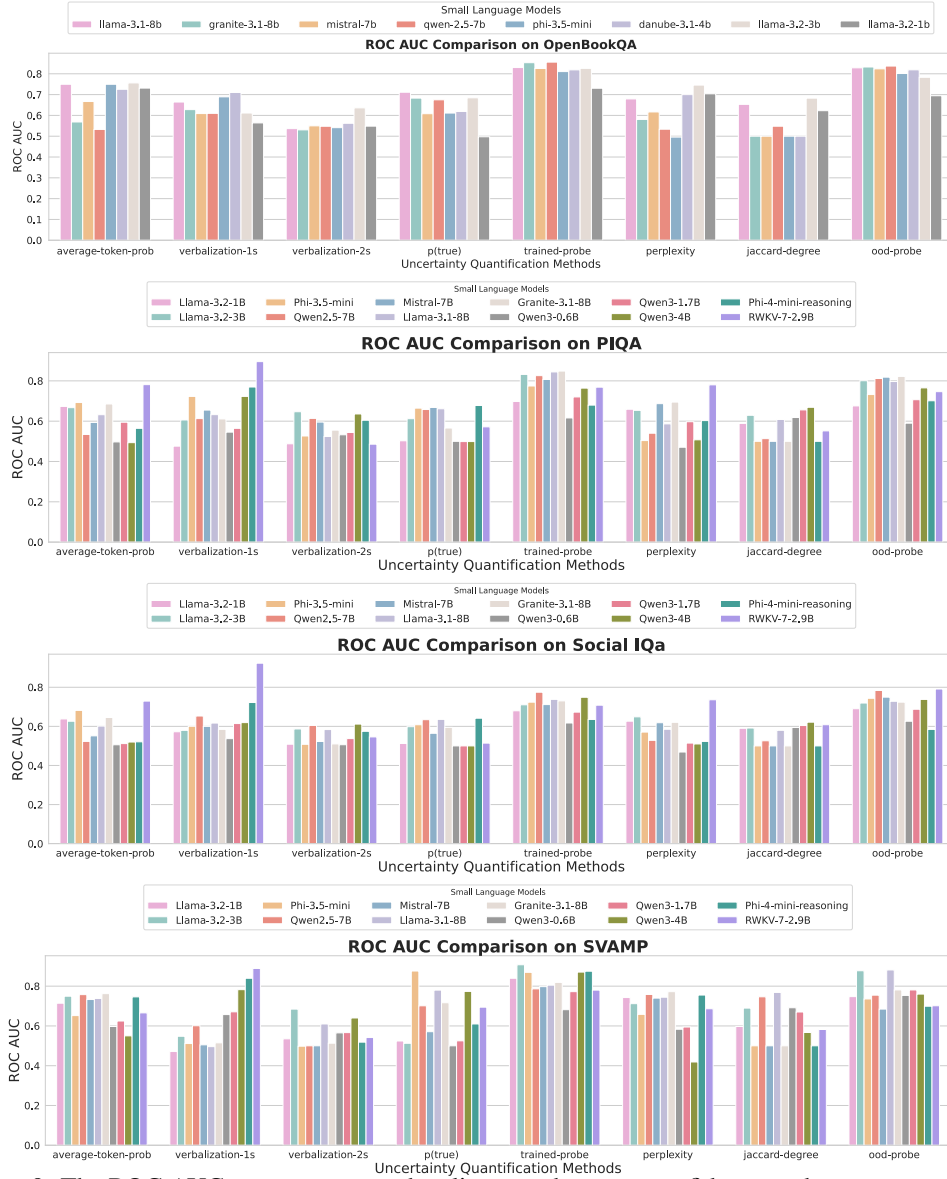


Figure 9: The ROC AUC scores measure the alignment between confidence and correctness across different SLMs and uncertainty quantification methods on OpenBookQA, PIQA, SocialIQA, and SVAMP datasets. A higher ROC AUC indicates a stronger alignment.

653 **Results of Alignment between uncertainty and correctness.**

654 All the experiments shown on this page are conducted under CommonsenseQA, SVAMP, TruthfulQA,
655 WinoGrande, and Math500 dataset with all 8 UQ methods.

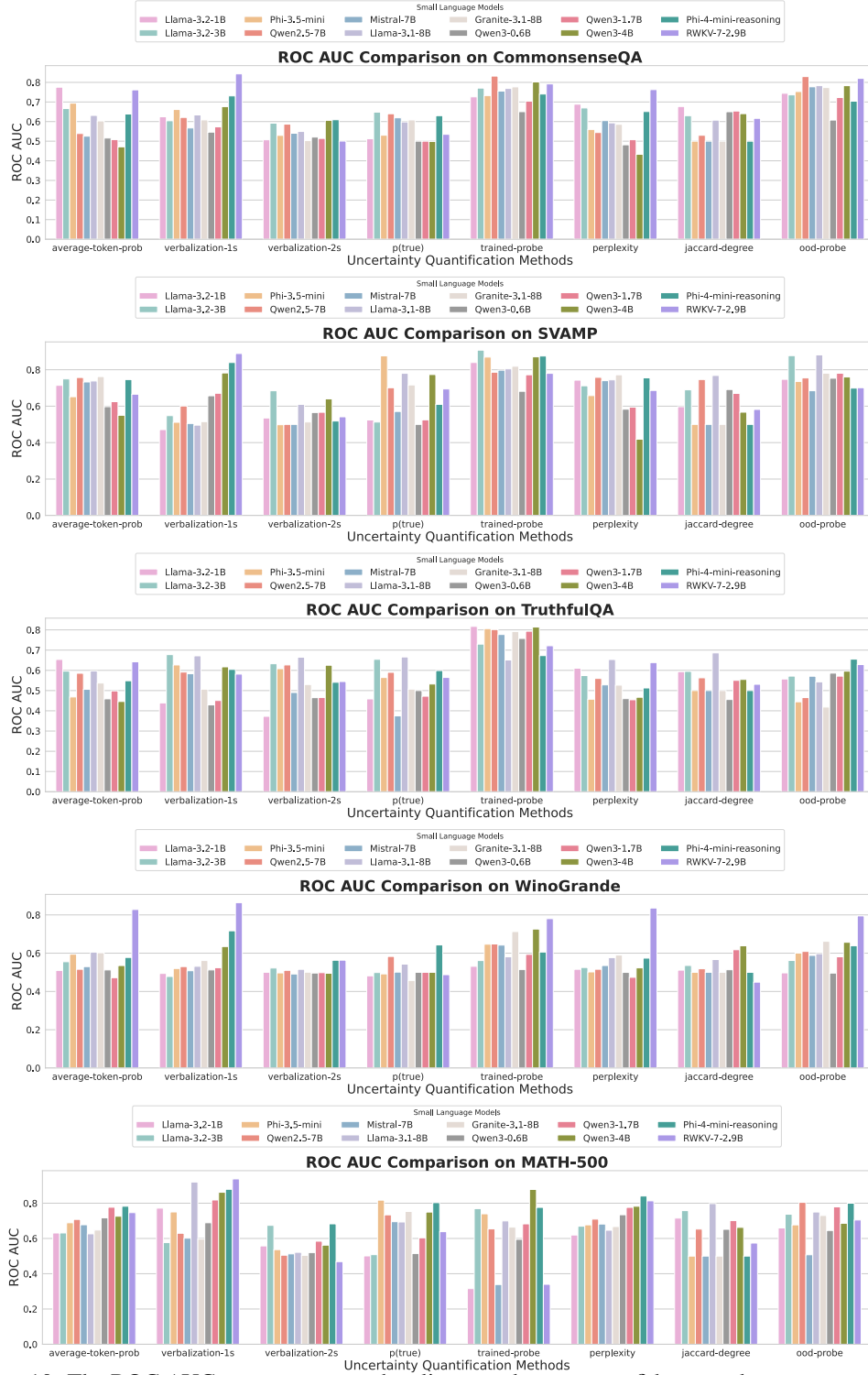


Figure 10: The ROC AUC scores measure the alignment between confidence and correctness across different SLMs and uncertainty quantification methods on CommonsenseQA. A higher ROC AUC indicates a stronger alignment.

656 C.2 Evaluation on Uncertainty-based Routing Approaches

657 Results of routing to GPT-4.1-Mini

658 All the experiments shown on this page are conducted under all benchmark datasets with selected
 659 SLMs. We only showcase partial of the experimental results.

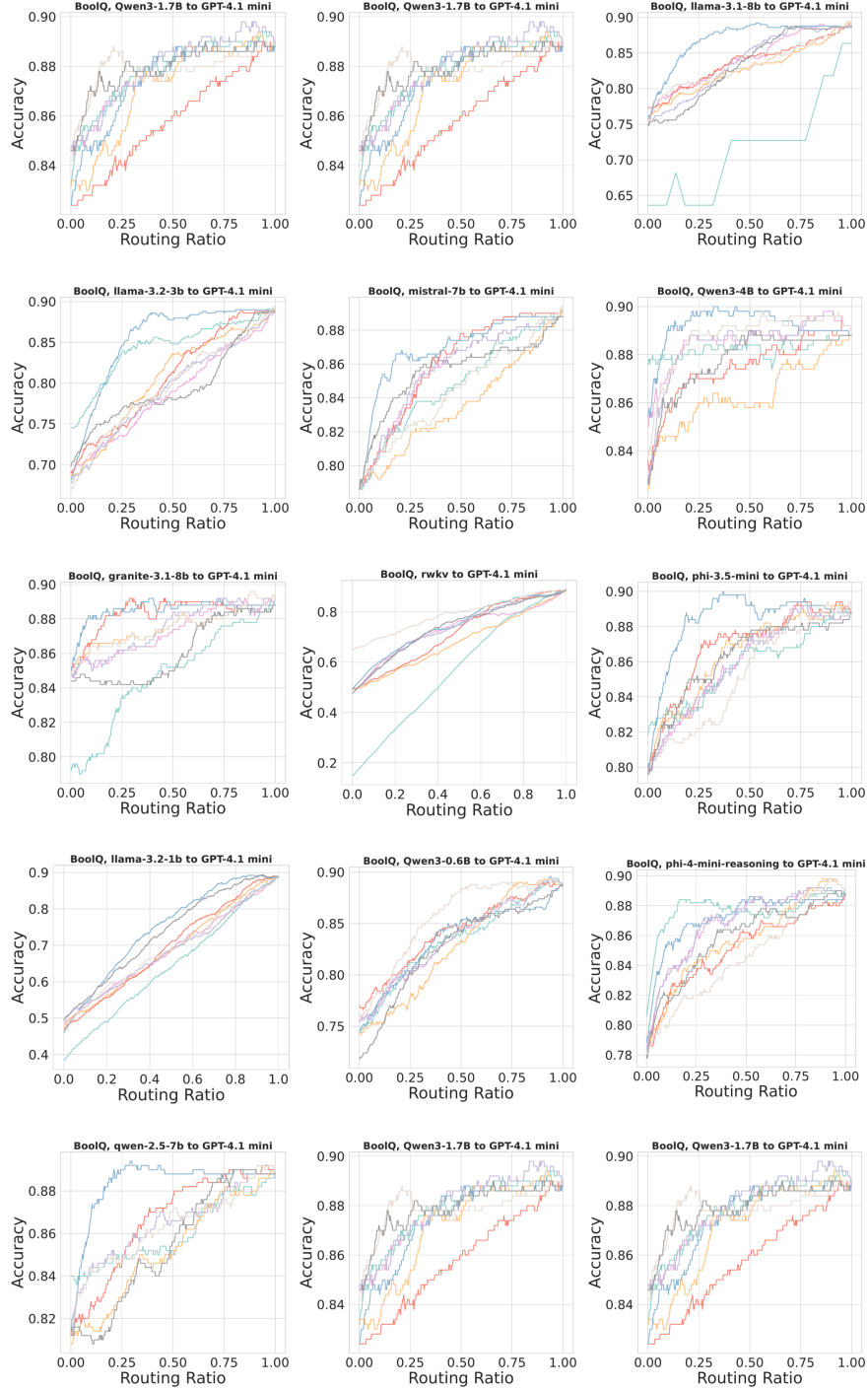


Figure 11: Overall accuracy vs. routing ratio with different UQ methods and SLMs.

660 Results of routing to DeepSeek-R1

661 All the experiments shown on this page are conducted under all benchmark datasets with selected
 662 SLMs. We only showcase partial of the experimental results.

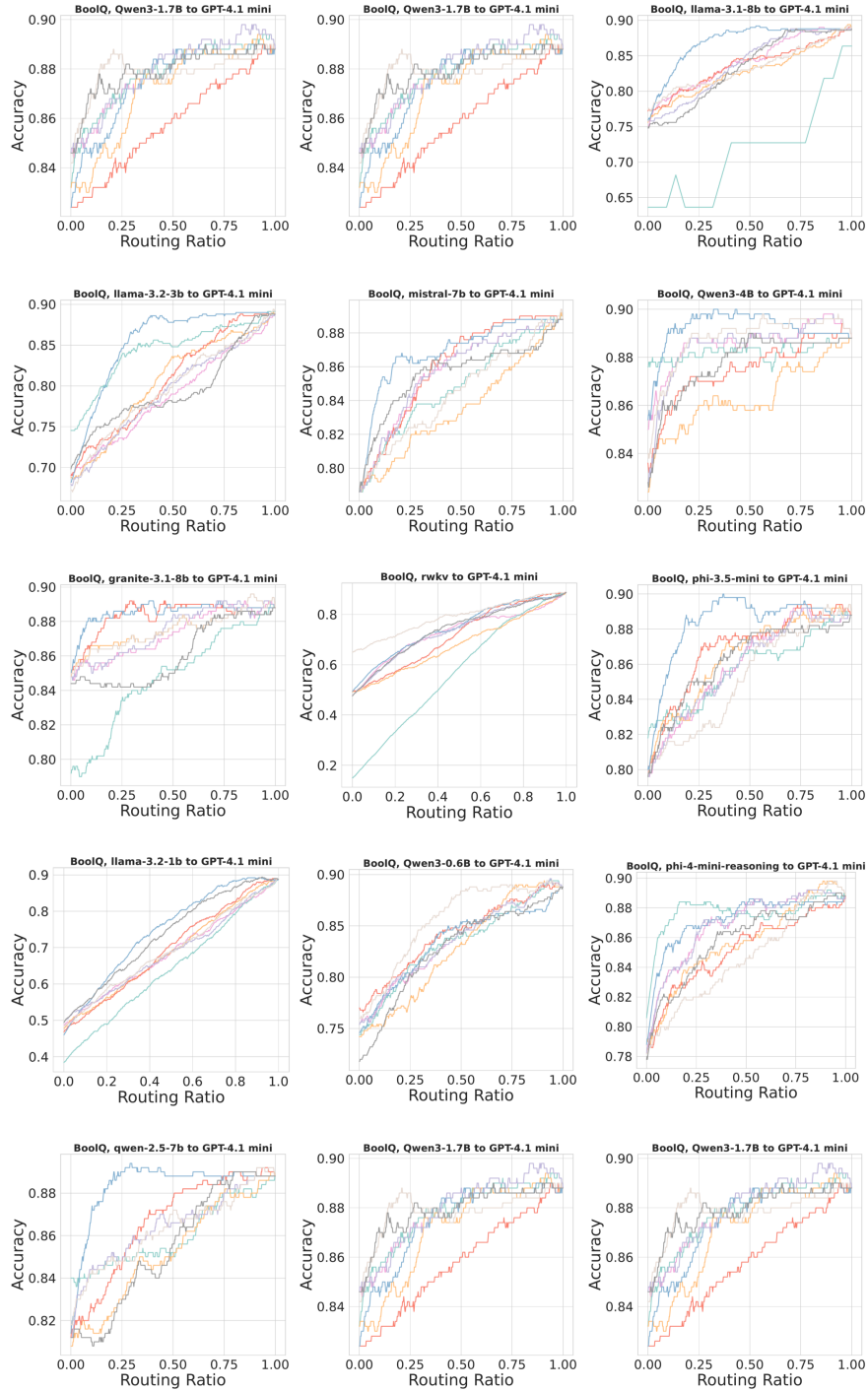


Figure 12: Overall accuracy vs. routing ratio with different UQ methods and SLMs.

663 **Results of routing to Llama-3.1-70B-Instruct**

664 All the experiments shown on this page are conducted under all benchmark datasets with selected
 665 SLMs. We only showcase partial of the experimental results.

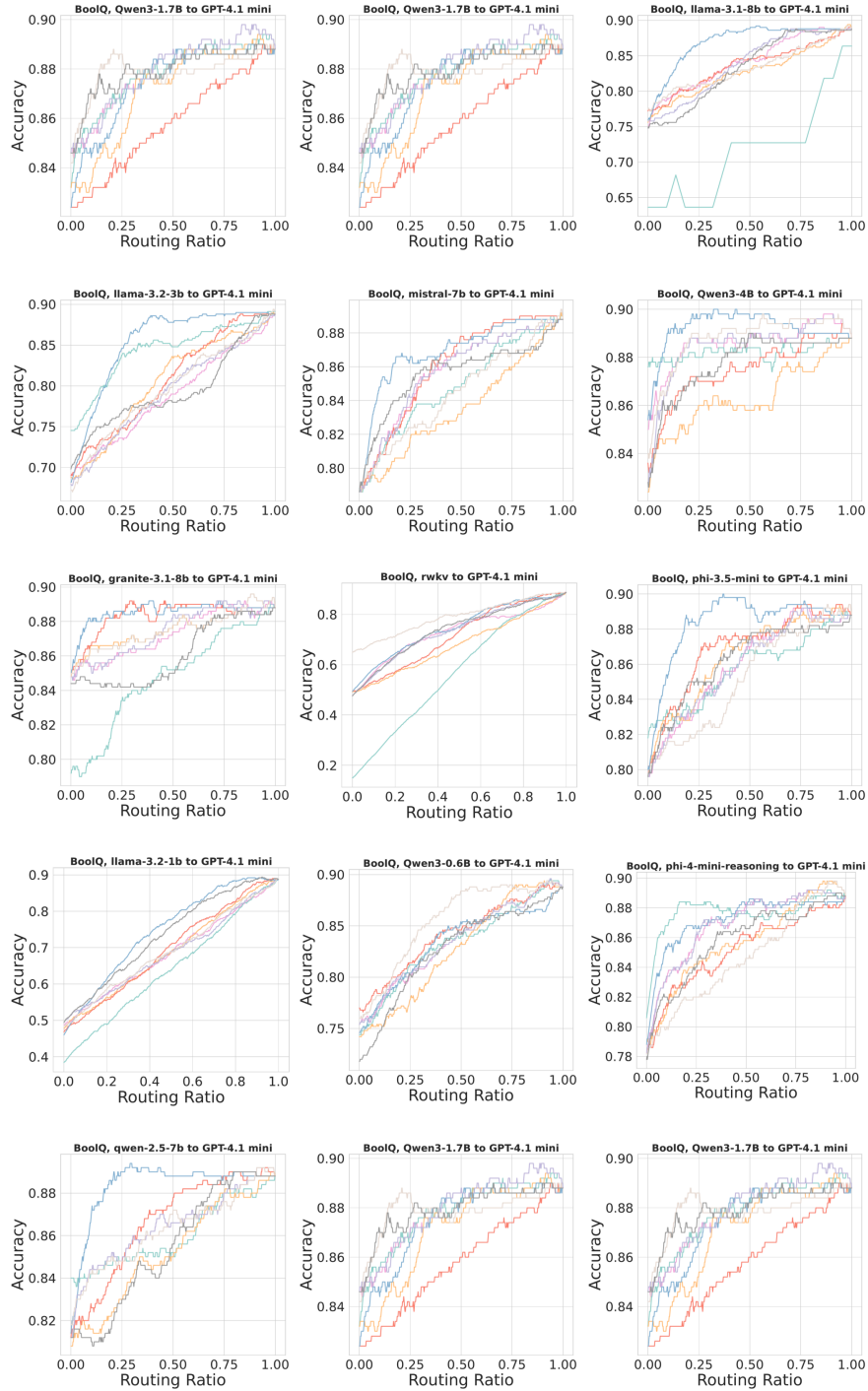


Figure 13: Overall accuracy vs. routing ratio with different UQ methods and SLMs.

666 Results of routing to Qwen3-32B

667 All the experiments shown on this page are conducted under all benchmark datasets with selected
 668 SLMs. We only showcase partial of the experimental results.

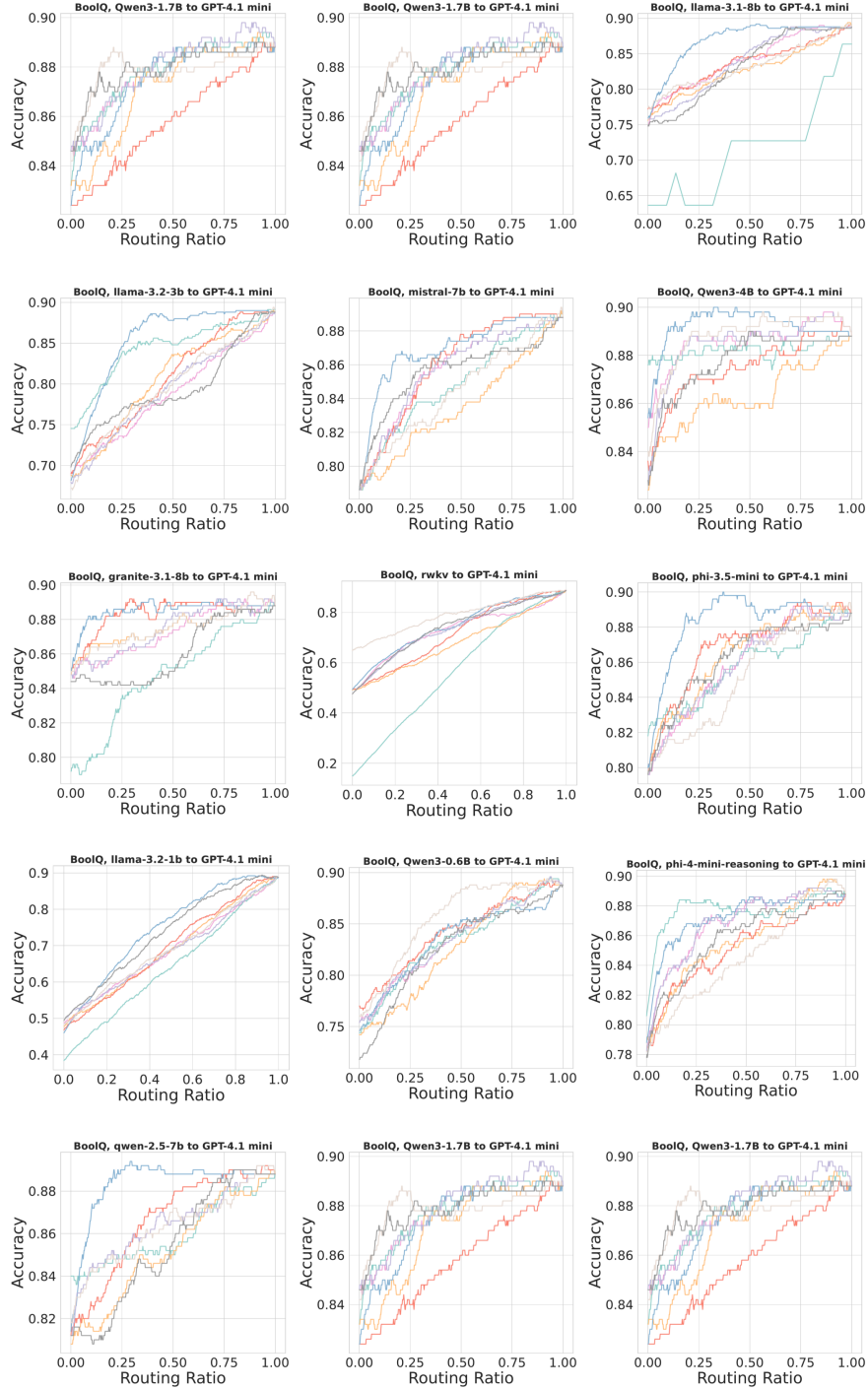


Figure 14: Overall accuracy vs. routing ratio with different UQ methods and SLMs.

669 C.3 Evaluation of Proxy Routing Data on New Downstream Scenario

670 Evaluation results on routing to GPT-4.1-Mini

The experiments shown on this page are conducted under all 15 datasets with different SLMs.

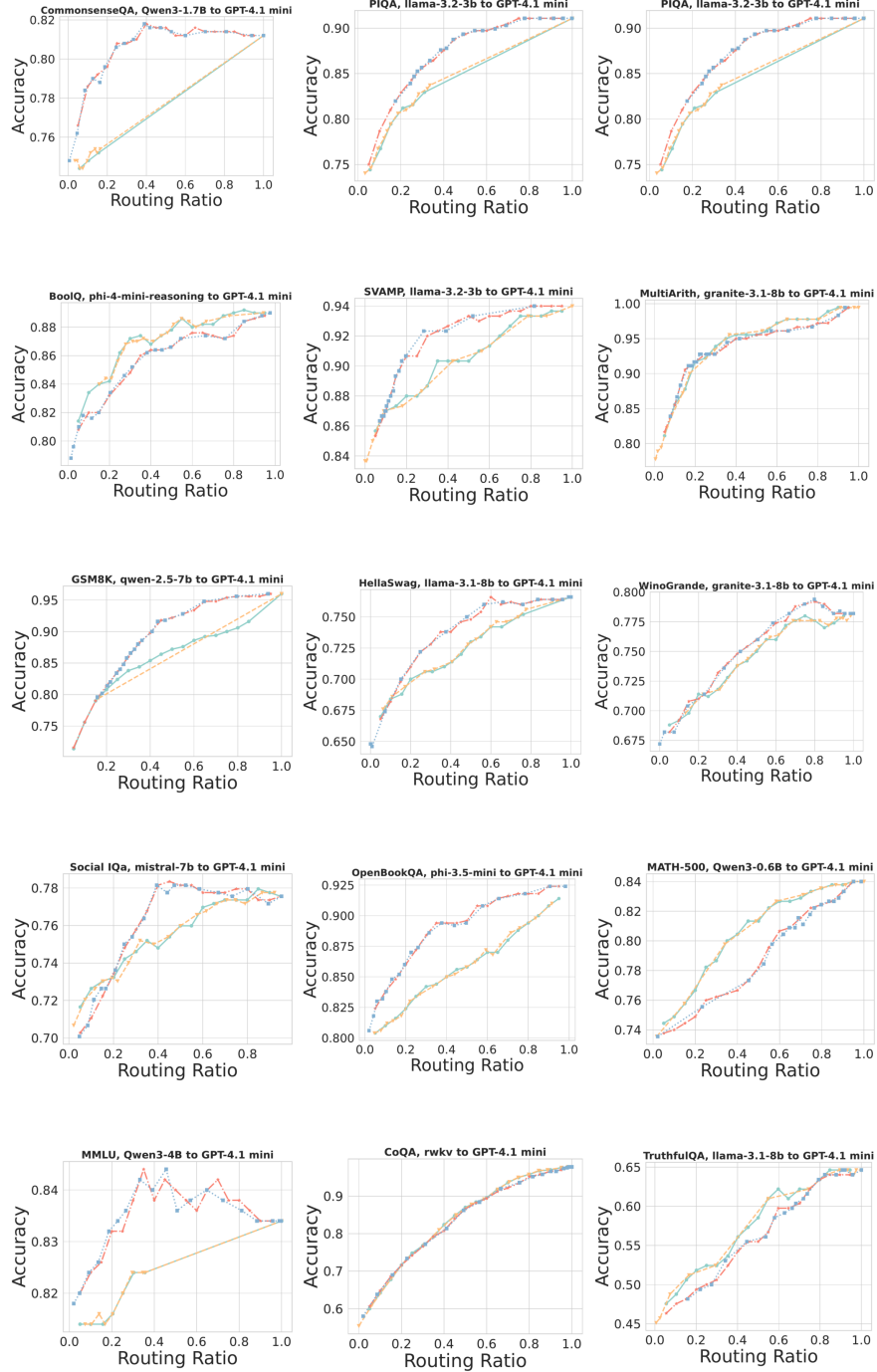


Figure 15: Assessing the generalization of proxy routing data to new downstream data for routing 12 SLMs to GPT-4.1-Mini on 15 datasets using two UQ methods (OOD Probe & Perplexity). The legend in Figure 5 is also used here.

672 Evaluation results on routing to DeepSeek-R1

The experiments shown are conducted under all math reasoning datasets with different SLMs.

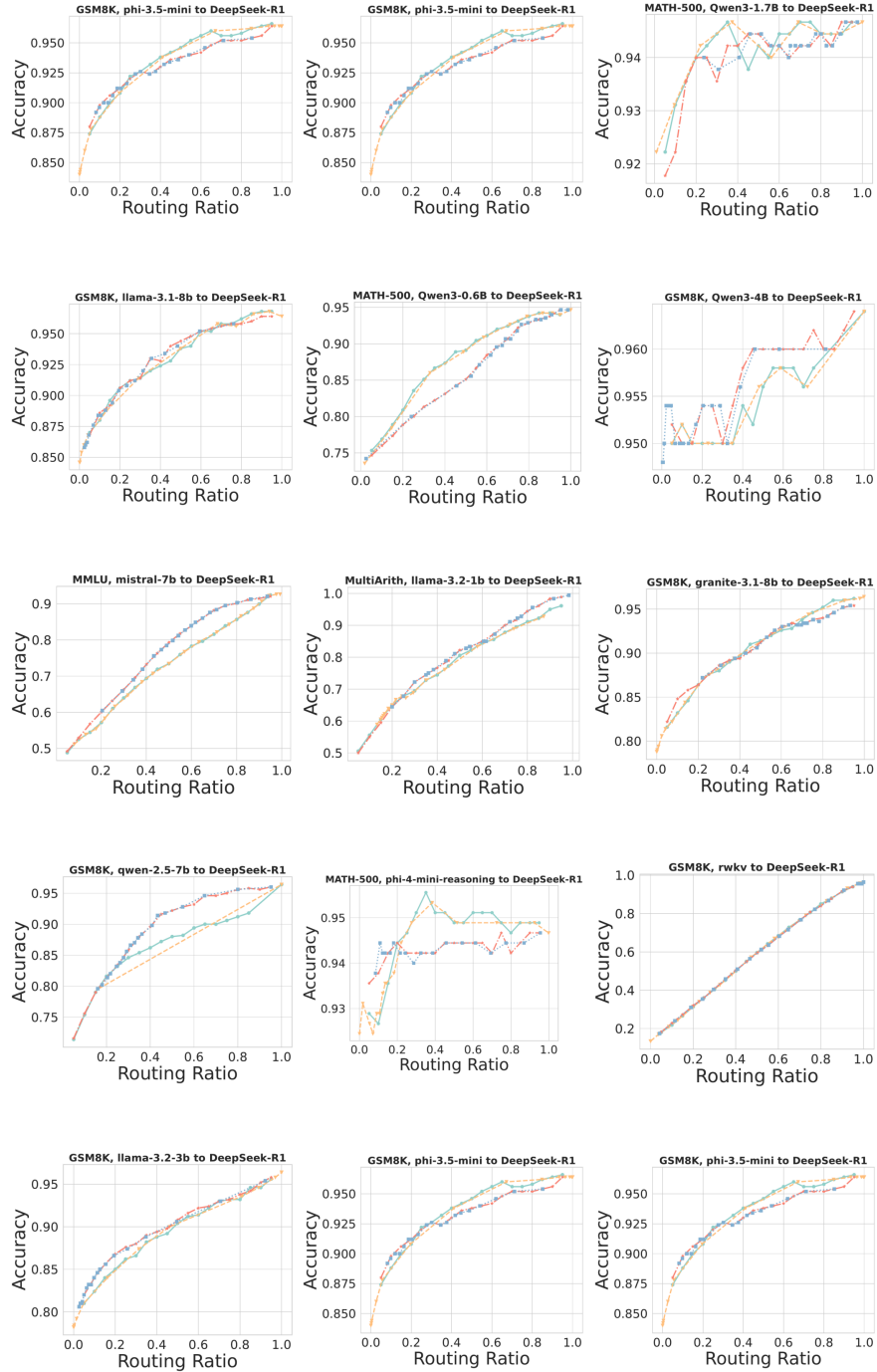


Figure 16: Assessing the generalization of proxy routing data to new downstream data for routing 12 SLMs to DeepSeek-R1 on 15 datasets using two UQ methods (OOD Probe & Perplexity). The legend in Figure 5 is also used here.

674 Evaluation results on routing to Llama-3.1-70B-Instruct

The experiments shown on this page are conducted under all 15 datasets with different SLMs.

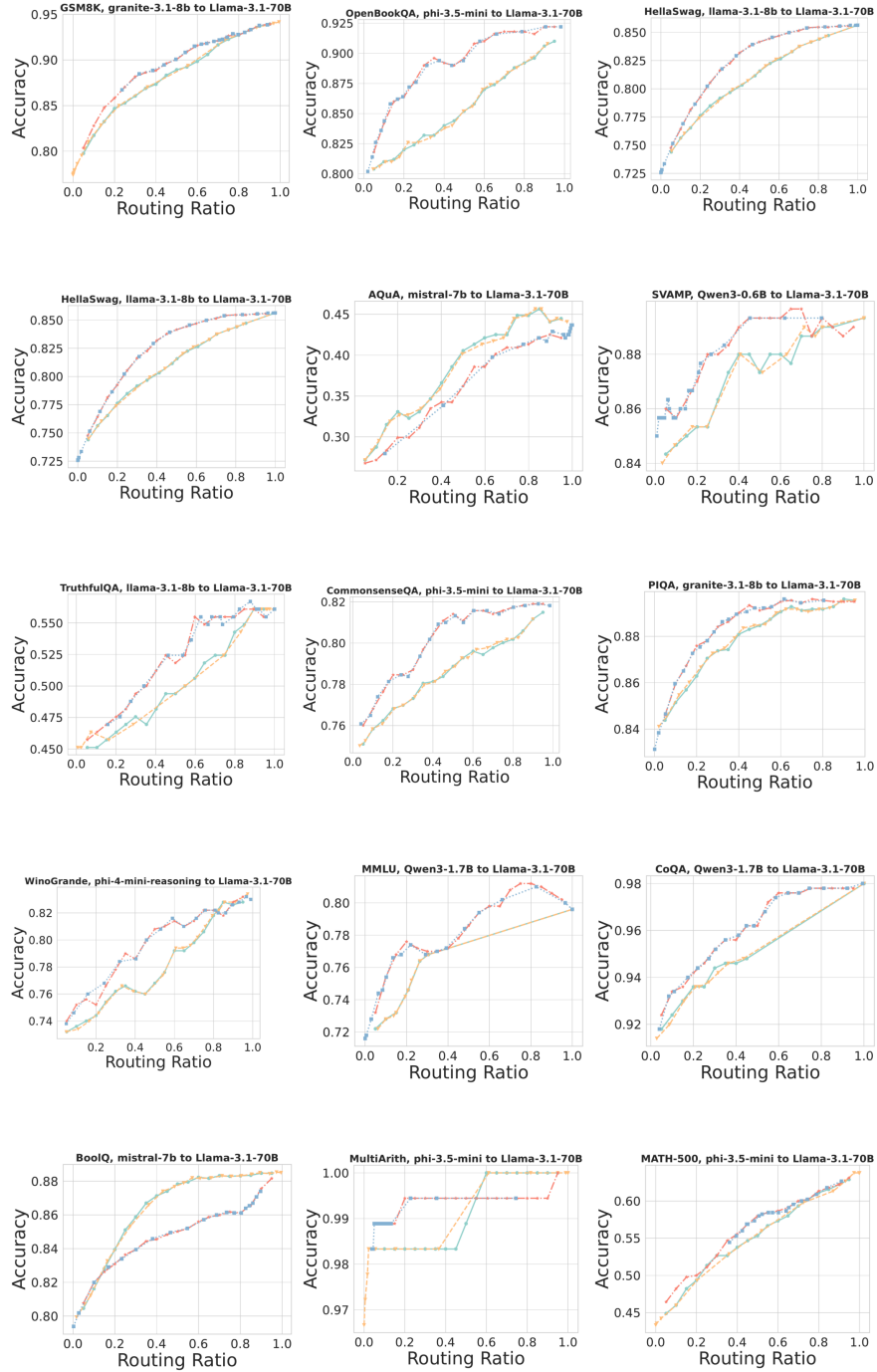


Figure 17: Assessing the generalization of proxy routing data to new downstream data for routing 12 SLMs to Llama-3.1-70B-Instruct on 15 datasets using two UQ methods (OOD Probe & Perplexity). The legend in Figure 5 is also used here.

676 **Evaluation results on routing to Qwen3-32B**

The experiments shown on this page are conducted under all 15 datasets with different SLMs.

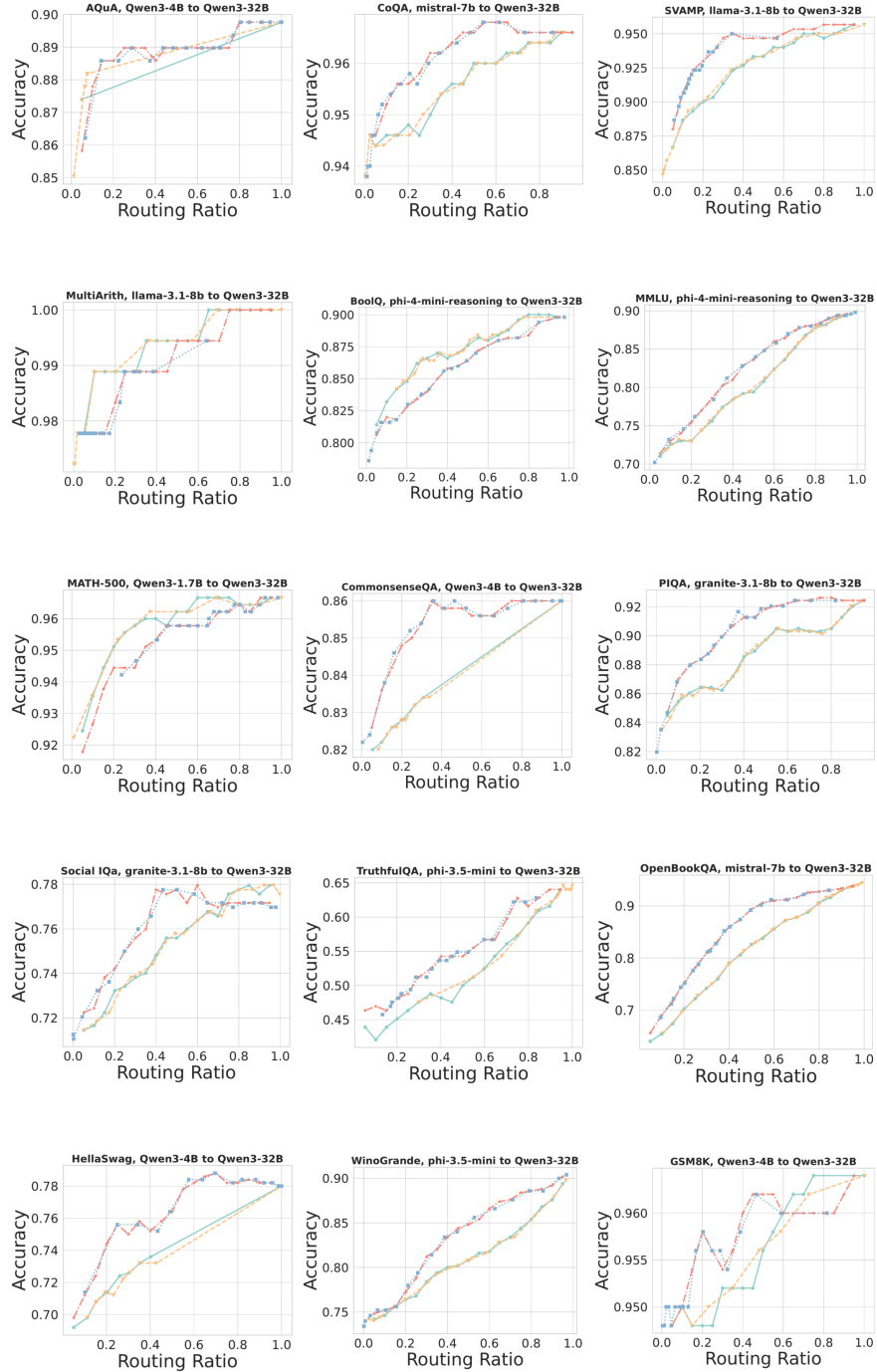


Figure 18: Assessing the generalization of proxy routing data to new downstream data for routing 12 SLMs to Qwen3-32B on 15 datasets using two UQ methods (OOD Probe & Perplexity). The legend in Figure 5 is also used here.

D Computational Infrastructure

The computational infrastructure information is given in Table 3.

Table 3: Experiment configuration and computing infrastructure.

Name	Value
Data type	torch.bfloat16
Flash-Attention	True
Computing Infrastructure	GPU
GPU Model	NVIDIA-A100
GPU Memory	80GB
GPU Number	4
CUDA Version	12.1
CPU Memory	512GB

E Limitation

While we do not foresee any immediate negative societal consequences, we acknowledge that uncertainty estimation is not infallible and that reliance on it may inadvertently introduce biases if the uncertainty scores are miscalibrated. Future work should consider evaluating the fairness and calibration of uncertainty estimates across diverse input populations.

F Impact Statement

The broader impact of this approach lies in its potential to democratize access to LLMs by significantly reducing computational cost, making high-quality language technologies more accessible to users with limited resources. This is particularly relevant in educational, humanitarian, and public-sector contexts where budget constraints often limit access to state-of-the-art AI models. Moreover, routing based on uncertainty has the potential to reduce over-reliance on any single model, thereby increasing robustness and resilience in real-world applications. This model-agnostic approach could encourage more modular and responsible deployment strategies in AI systems.

G Routing with Proxy Routing Data vs. Random Routing

Our approach sets a confidence threshold for routing on a new dataset without prior access, and no existing work has addressed this scenario to the best of knowledge. To quantify its effectiveness we compare it with a *random routing* baseline. On HellaSwag, we route three SLMs to GPT-4.1 mini using Perplexity. We compute the average root mean squared (RMS) distance between the oracle routing curve derived from the full downstream dataset and those obtained using either our proposed method or random routing. A lower RMS distance indicates closer alignment with the oracle and therefore better routing quality. As shown in Table 4, routing based on proxy routing data dramatically outperforms random routing. Similar gains are observed with other uncertainty quantification methods (e.g., a 34.14% improvement with OOD Probe).

Table 4: Routing with proxy routing data vs. random routing on HellaSwag. Lower RMS is better.

Method	LLAMA-3.2-3B	MISTRAL-7B	LLAMA-3.1-8B
Ours	0.001	0.001	0.001
Random routing	0.029	0.031	0.019

Robustness under strong OOD shifts. We further test our approach based on proxy routing data in two challenging out-of-distribution (OOD) scenarios:

- **Math** \rightarrow **Commonsense**: Proxy routing data drawn solely from math datasets (GSM8K, AQuA, MultiArith, SVAMP); evaluation on commonsense reasoning (TruthfulQA), routing various SLMs to Llama-3.1-70B.
- **Commonsense** \rightarrow **Math**: Proxy data drawn solely from commonsense datasets; evaluation on the math dataset AQuA with the same routing setup.

Table 5: Routing with proxy routing data vs. random routing under strong OOD shifts.

OOD Setting	Method	PHI-3.5-MINI	MISTRAL-7B	LLAMA-3.1-8B
Math \rightarrow Commonsense	Ours	0.0148	0.0048	0.0132
	Random routing	0.0187	0.0090	0.0176
Commonsense \rightarrow Math	Ours	0.0057	0.0060	0.0022
	Random routing	0.0085	0.0082	0.0037

Across both OOD shifts (Table 5), routing with proxy routing data consistently yields smaller RMS distances than the random baseline, underscoring its strong generalization capability.