No Task Is an Island: A Survey of Low-Resource Task Adaptation in LLMs

Anonymous ACL submission

Abstract

Task adaptation is a crucial step for deploying large language models (LLMs) on downstream tasks. However, in low-resource settings, the scarcity of labeled data limits effective task-specific specialization, often leading to performance degradation, hallucinations, and unreliable behavior in long-tail applications. In this paper, we present a systematic survey of low-resource task adaptation and propose a taxonomy spanning data-centric methods, effective post-training, and reliable inference. We review representative approaches, summarize resources, and highlight open challenges. These discussions frame low-resource task adaptation as an important research problem for reliable LLM deployment and as part of a broader shift toward resource-aware NLP research.

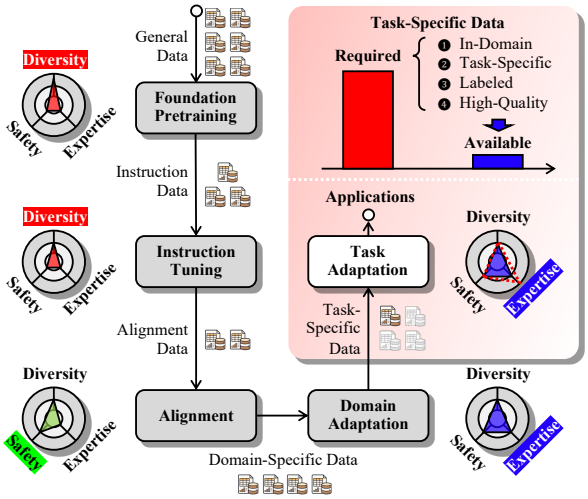

Figure 1: Pipeline for downstream LLM deployment.

1 Introduction

"No man is an island, entire of itself; every man is a piece of the continent, a part of the main."
— JOHN DONNE,
Devotions upon Emergent Occasion (1624)

Recently, large language models (LLMs) have shown remarkable performance across diverse NLP tasks. Trained on large-scale general corpora, foundation models learn broadly transferable representations (Yang et al., 2025a; Brown et al., 2020; Touvron et al., 2023), and subsequent instruction tuning and alignment further improve their usability and safety (Wang et al., 2023d; Rafailov et al., 2023; Ouyang et al., 2022). Despite these advances, achieving optimal performance on target applications often requires additional specialization through model adaptation (Gururangan et al., 2020; Howard and Ruder, 2018). In addition to domain adaptation, task-specific supervision is widely used to tailor model behavior to downstream objectives, a practice commonly referred to as task adaptation.

In particular, task adaptation can be viewed as the process of enabling foundation models to acquire target capabilities from task-specific data. This process is typically implemented as a post-training step via supervised learning (Lingam et al., 2024; Hu et al., 2022; Li and Liang, 2021), which updates model parameters using as much labeled data as is available. The resulting model is then applied at inference time, where in-context learning and reasoning strategies can be incorporated to improve reliability and performance (Borgeaud et al., 2022; Wei et al., 2022b; Lewis et al., 2020).

The effectiveness of task adaptation heavily depends on the availability of high-quality supervision for the target task. However, labeled data for downstream objectives are often scarce because of costly annotation and strict data requirements in practice (see Figure 1). Such data scarcity exists in a wide range of LLM-based systems (Zhang et al., 2024a; Wang et al., 2023c; Wei et al., 2022a), although the resulting performance loss appears less visible by using stronger foundation models and more aggressive domain adaptation. Nevertheless, Isik et al. (2025) observe that such compensation is not consistently reliable across different evaluation metrics. Additionally, the rapid explosion of new

tasks, such as emerging programming languages, further amplifies the urgency of effective task adaptation methods under low-resource settings.

Regarding data scarcity in LLMs, existing surveys primarily focus on low-resource human languages (Qin et al., 2025a; Zhu et al., 2024a), or on specific methodological threads, such as data annotation and synthesis (Luo et al., 2025b; Tan et al., 2024), the use of external knowledge (Dong et al., 2024; Li et al., 2025d), and knowledge boundaries of LLMs (Wang et al., 2024c; Li et al., 2025b). While these works conduct comprehensive reviews either of the limitations of pretraining corpora or of individual research aspects, they do not offer a holistic perspective on low-resource task adaptation, in which the key bottleneck is the scarcity of high-quality and task-specific supervision.

This survey presents a systematic review of task adaptation, with an emphasis on robust deployment with limited supervision. We propose a stage-based taxonomy (§2), spanning data-centric methods, effective post-training, and reliable inference, aiming to answer three key questions: ❶ How can we obtain and leverage high-quality supervision? (§3); ❷ How can we perform effective model updating under limited labeled data? (§4); ❸ How can we improve reliability at inference time for low-resource LLMs? (§5). Finally, we summarize open challenges (§6) and organize available resources and forward-looking perspectives (§A), highlighting low-resource task adaptation as a new mission for NLP research: moving beyond general-purpose benchmarks toward resource-aware deployment for underrepresented tasks and communities.

2 Preliminary and Taxonomy

In this paper, we focus on task adaptation under limited supervision. We adopt a broader definition of task adaptation that includes both parameter updating and inference, as the latter also contributes significantly to the issues caused by data scarcity.

Under scarce supervision, it is common practice to aggregate training signals from multiple sources (e.g., datasets from related tasks). Given seed labeled datasets $\{D_S^1, ..., D_S^n\}$ from n tasks, the datasets applied at the t_{th} post-training step are:

$$\{D_A^{1,t}, ..., D_A^{n,t}\} = \chi(\{D_S^1, ..., D_S^n\}; M^{t-1}, \mathbb{D}_R)$$

where M^{t-1} refers to the model parameters from the last training step, $\mathbb{D}_R$ denotes the set of auxiliary unlabeled data resources, and $\chi(\cdot)$ denotes

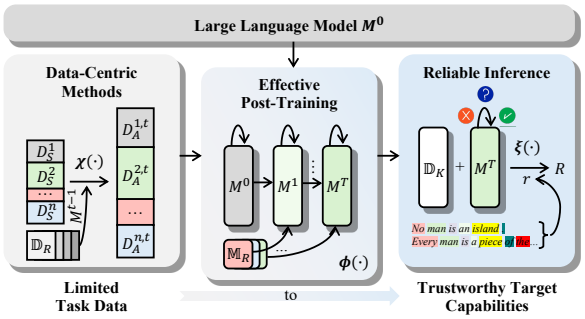

Figure 2: An overview of low-resource task adaptation.

the data-centric procedure that explores more reliable supervision and performs suitable data mixing across training stages. Based on the constructed task data, effective post-training is applied to adapt the LLM parameters to the target task. The parameter set at step t is updated as:

$$M^t = \phi(\{D_A^{1,t}, ..., D_A^{n,t}\}, M^{t-1}; \{M^0, ..., M^{t-2}\}, \mathbb{M}_R)$$

where $\phi(\cdot)$ denotes the post-training function of task adaptation. To further enhance post-training, $\phi(\cdot)$ can optionally leverage earlier checkpoints $\{M^0, ..., M^{t-2}\}$ and the set of auxiliary models $\mathbb{M}_R$ to extract training dynamics, model signals, parameter distributions, and prior knowledge for faster convergence or better performance.

After iterative parameter updates, the adapted LLM M^T is ready for inference. Given a query Q, the adapted model generates its response as:

$$R = \xi(Q, M^T, r; \mathbb{D}_K)$$

where $\xi(\cdot)$ denotes the inference process. Notably, the adapted model M^T can be fragile in downstream use under low-resource supervision. To improve the reliability, an external knowledge or tool base $\mathbb{D}_K$ and a flexible inference strategy r can be incorporated at test time to support reasoning and ground the model's responses.

An overview of low-resource task adaptation is shown in Figure 2. In this survey, we organize the related works in three categories: data-centric methods, effective post-training, and reliable inference. We primarily focus on challenges and the corresponding approaches strongly related to data scarcity, but also include the supporting studies that are highly relevant[1]. Figure 3 presents a taxonomy of low-resource task adaptation, and detailed discussions are provided in the following sections.

[1]Appendix A.1 discusses the scope of our survey in detail.

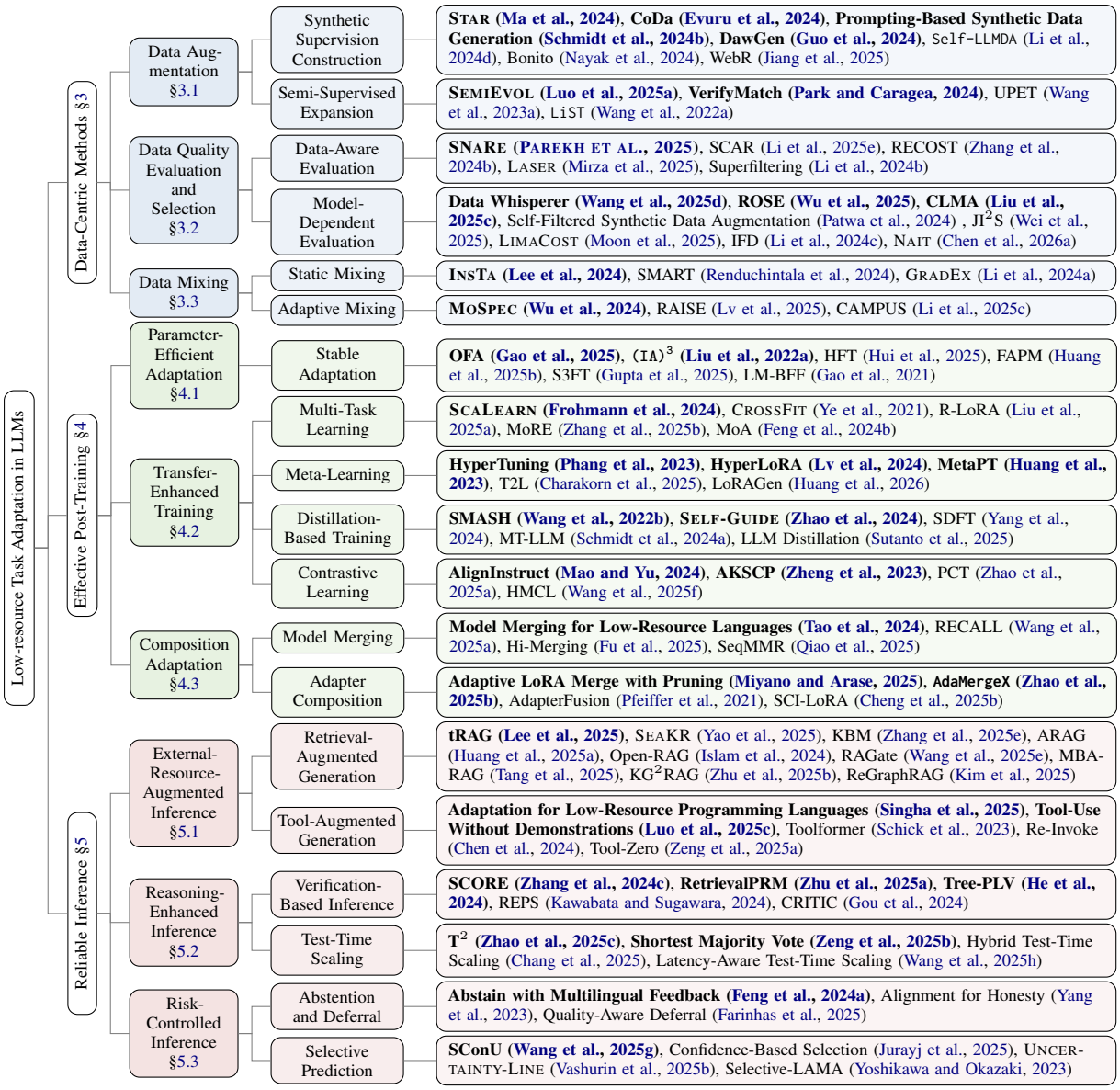

Figure 3: A taxonomy of low-resource task adaptation in LLMs. Boldfaced papers indicate direct low-resource adaptation studies; others are essential supporting works. Detailed comparisons are provided in Appendix A.1.

3 Data-Centric Methods

Data scarcity is a fundamental challenge in low-resource task adaptation. To address this, data-centric methods aim to improve both the quantity and quality of supervision. These approaches can be viewed as different designs of the data-centric procedure $\chi(\cdot)$, which governs how supervision is expanded and mixed across training stages.

3.1 Data Augmentation

Data augmentation addresses the low-resource issue by constructing pseudo training data to expand task-relevant supervision. In this survey, we distinguish between synthetic supervision construction, which directly generates new supervision using generative models, and semi-supervised expansion, which leverages real unlabeled data to produce labeled task-specific samples.

Synthetic Supervision Construction. A central question in synthetic supervision construction is how to generate task data while ensuring compatibility with downstream task requirements. Existing studies find it is beneficial to break data construction into simpler subproblems. For example, Ma et al. (2024) first generate target structures and then produce corresponding passages in low-resource information extraction, and Schmidt et al. (2024b) sample answer candidates from contexts and generate questions conditioned on them for question answering. Additionally, researchers emphasize that naively generated pseudo-data may deviate sub-

stantially from the real few-shot data distribution. Evuru et al. (2024) address this issue by extracting simple instance-level constraints and incorporating them into prompts for controllable generation, whereas Guo et al. (2024) explicitly align the synthetic data generator with the target few-shot distribution. Beyond these task-oriented methods, the work on instruction data construction also provides perspectives for synthetic supervision (Li et al., 2024d; Nayak et al., 2024; Nguyen et al., 2024).

Semi-Supervised Expansion. Semi-supervised expansion focuses on how to exploit abundant unlabeled data to enlarge task supervision. This problem has been systematically studied since the era of pretrained language models (Wang et al., 2022a, 2023a). More recently, for LLM task adaptation, Luo et al. (2025a) introduce a propagate-and-select framework that first transfers knowledge from labeled data to unlabeled data with in-weight and in-context propagation, and then selects higher-quality pseudo-responses via collaborative learning. Park and Caragea (2024) further highlight the importance of verifying the reliability of augmented data during semi-supervised learning. Instead of simply discarding low-confidence samples, Park and Caragea (2024) improve their utilization through confidence-aware MixUp with human-labeled data.

3.2 Data Quality Evaluation and Selection

The quality of both real and synthetic task-specific data largely determines the performance of adapted LLMs. In this survey, we consider data-aware and model-dependent evaluation approaches.

Data-Aware Evaluation. Data-aware evaluation studies how to estimate task-relevant instances using sample-level signals. Related methods are also applicable in instruction tuning and pre-training. For data-aware evaluation, Li et al. (2025e) show that stylistic consistency of training responses is an important factor affecting tuning effectiveness. Zhang et al. (2024b) further introduce external knowledge to assess synthetic data and combine knowledge-based re-ranking to better identify reliable samples. Mirza et al. (2025) propose a unified selection pipeline that jointly considers data quality, difficulty, and diversity, while Li et al. (2024b) explore whether lightweight proxy models can effectively replace stronger models in data filtering. In addition, Parekh et al. (2025) suggest that in low-resource adaptation, low-quality synthetic samples should be refined rather than merely identified.

Model-Dependent Evaluation. Model-dependent evaluation emphasizes that the value of task data is closely tied to model-related signals during training and inference. Wu et al. (2025), Moon et al. (2025), Wei et al. (2025), and Mekala et al. (2024) estimate sample utility through optimization-related signals (e.g., preference loss, gradient- or influence-based valuation, and training dynamics). Wang et al. (2025d) further explore a training-free way, where the target model scores candidate samples through in-context learning. Chen et al. (2026a) move to an even finer granularity by selecting samples according to neuron activation features that are more relevant to the target task capability. Meanwhile, Liu et al. (2025c) and Patwa et al. (2024) extend model-dependent evaluation to pseudo-data construction, respectively via collaborative pseudo-annotation and in-context self-filtering.

3.3 Data Mixing

Data mixing addresses data scarcity from the perspective of data composition to enrich task-relevant supervision. Although most existing mixing strategies are designed for instruction tuning or developing multi-task models, related work offers transferable insights for low-resource task adaptation.

Static Mixing. Static mixing studies how to determine a fixed data composition before training. Longpre et al. (2023) and Xie et al. (2023) demonstrate that mixture design can substantially affect convergence and final performance. For adaptation to specific tasks, Iyer et al. (2024) and Lee et al. (2024) emphasize that mixing irrelevant tasks may hurt target-task performance and Lee et al. (2024) propose instruction-based task selection for a more effective mixture. Renduchintala et al. (2024) and Li et al. (2024a) consider static mixing as an optimization problem, focusing on task allocation and efficient subset selection. Beyond mixture proportions, Rao et al. (2024) further explores how the organization of fixed data partitions during training can improve the effectiveness of data mixing.

Adaptive Mixing. Unlike static mixing, adaptive mixing focuses on dynamically adjusting data usage according to the model state during training. Existing studies investigate this line at different granularities: Wu et al. (2024) propose MoS and explore dataset-level adaptive reweighting and sampling, while Lv et al. (2025) focus on instance-level adaptive selection. Notably, based on MoS, Wu et al. (2024) further propose a task-specific variant, MoSpec, to support adaptation to particular target tasks. Beyond dataset- and instance-level adjust-

ment, Li et al. (2025c) study adaptive mixing from a curriculum-learning perspective, dynamically organizing sub-curricula and training schedules with difficulty views and competence-aware scheduling.

> ***Takeaway: How can we obtain and leverage high-quality supervision?***
> §3.1: Expanding the **quantity** of task-specific supervision by constructing pseudo-data from limited labeled task data D_S^1 and abundant unlabeled data $\mathbb{D}_R$ with model M^{t-1};
> §3.2: Improving the supervision **quality** by identifying reliable instances within data-centric procedures $\chi(\cdot)$;
> §3.3: Broadening the **scope** of supervision by incorporating additional data sources $\{D_S^2, ..., D_S^n\}$, optionally with their usage adjusted based on the model state from M^{t-1}.

4 Effective Post-Training

Limited labeled data tends to make task adaptation unreliable or even lead to training failure. For this issue, effective post-training methods are proposed to better utilize scarce task-specific supervision with robust optimization. In this survey, methods of this kind are viewed as efforts to improve the post-training process $\phi(\cdot)$ under low-resource settings.

4.1 Parameter-Efficient Adaptation

Parameter-efficient fine-tuning (PEFT) has been widely adopted in post-training due to its efficiency in computation and memory footprint (Liu et al., 2022b; Hu et al., 2022; Zhang et al., 2023a). For low-resource task adaptation, PEFT is particularly appealing because restricting the number of trainable parameters is helpful in mitigating overfitting (He et al., 2021; Shuttleworth et al., 2025). However, how to maintain a stable adaptation under scarce supervised data remains a critical challenge.

Stable Adaptation. Stable parameter-efficient adaptation aims to maintain robust and reliable post-training with limited supervision. Gao et al. (2025) revisit and improve few-shot adaptation with an optimization-inspired strategy. Liu et al. (2022a) instead adapt LLMs by learning only a small set of scaling vectors on internal activations, yielding a more stable adaptation within a highly constrained update space. Gupta et al. (2025) attribute instability partly to the single-gold-response assumption, and alleviate this issue by incorporating the model responses as additional supervision. Meanwhile, Hui et al. (2025) and Huang et al. (2025b) address catastrophic forgetting via selective parameter freezing and pruning of task vectors.

4.2 Transfer-Enhanced Training

Due to the scarcity of task-specific data, transfer-enhanced methods are explored to improve post-training by introducing transferable supervision. These methods are broadly from four perspectives: leveraging related tasks, learning from prior optimization experience, distilling knowledge from other models, and constructing auxiliary supervision from alternative views of training instances.

Multi-Task Learning. Before the LLM era, Ye et al. (2021) show that upstream learning over diverse tasks can improve adaptation to unseen few-shot tasks. Building on this transfer view, Pfeiffer et al. (2021) propose a two-stage transfer framework for adapter-based pretrained models. Frohmann et al. (2024) further extend this paradigm to LLMs by enabling task transfer via lightweight parameter scaling over task knowledge. Beyond transfer effectiveness, task interference remains another challenge in multi-task adaptation. Feng et al. (2024b) and Zhang et al. (2025b) formulate multi-task adaptation as a mixture of LoRA experts, using explicit routing or adaptive rank selection to mitigate task conflicts. Liu et al. (2025a), in contrast, emphasize the importance of capturing task-specific differences, and reduce interference with diversified multi-head LoRA parameterization.

Meta-Learning. Meta-learning aims to acquire transferable adaptation behavior that supports fast and effective learning under scarce supervision. Early studies such as Bansal et al. (2022) and Hong and Jang (2022) apply this idea to adapter-based tuning by learning few-shot-friendly adapter spaces. Phang et al. (2023) further reframe meta-learning as parameter generation, using a hypernetwork to directly produce task-specific PEFT parameters. More recent work extends this paradigm to LoRA-style adaptation. For example, Lv et al. (2024) improve hypernetwork stability in low-resource settings, Charakorn et al. (2025) enable instant one-pass adaptation from short task descriptions, and Huang et al. (2026) emphasize the structural properties of LoRA parameter spaces. In parallel, Huang et al. (2023) show that meta-learning can also benefit prompt tuning by learning more reliable soft prompt initialization when task data are scarce.

Distillation-Based Training. Distillation-based training addresses the low-resource issue by transferring knowledge from teacher models to compensate for scarce supervision. Wang et al. (2022b) improve few-shot adaptation by intermediate-task-based distillation, where collaborative design between tasks and prompts enables more transferable knowledge. Zhao et al. (2024) and Yang et al. (2024) further explore self-distillation. The former

synthesizes task-specific data for subsequent fine-tuning, while the latter aligns task supervision with the seed model distribution to reduce catastrophic forgetting. Besides, Schmidt et al. (2024a) and Sutanto et al. (2025) extend distillation-based training to more specialized scenarios, such as low-resource cross-lingual transfer and multiple-choice QA.

Contrastive Learning. Contrastive learning enhances low-resource adaptation by constructing auxiliary supervision from limited data. Such contrastive supervision can be instantiated diversely across tasks. Mao and Yu (2024) strengthen low-resource machine translation via word-alignment-based contrastive supervision, while Zheng et al. (2023) target low-resource NLU by combining cloze-driven prompting with contrastive learning to better activate pretrained knowledge and improve generalization. Zhao et al. (2025a) instead focus on the granularity of optimization, decomposing prompt embeddings and applying contrastive learning only to the task-specific vectors for more precise transfer. More generally, Wang et al. (2025f) show that hierarchical multi-level contrast benefits complex tasks, especially in low-resource settings.

4.3 Composition Adaptation

Unlike transfer-enhanced training, composition adaptation addresses low-resource task adaptation by combining knowledge from multiple models or task-specific parameter modules when a single model or adaptation path is insufficient.

Model Merging. Model merging improves post-trained models by combining existing models with little or no training. In low-resource settings, Tao et al. (2024) and Sarasua et al. (2025) find that adapting a single model may fail to balance language and task-specific abilities, and propose to separate these adaptation goals into various models and merge them afterward. Another central challenge in model merging is interference between merged models. Wang et al. (2025a) address this issue through adaptive hierarchical weighted fusion based on layer-wise representation similarity, while Fu et al. (2025) further reduce parameter conflicts with scaling and pruning. Qiao et al. (2025) extend model merging to sequential knowledge editing, suggesting its broader potential for stabilizing models across different training stages.

Adapter Composition. We highlight adapter composition due to the widespread adoption of adapter-based methods. Early work shows that first learning task-specific adapters and then composing them can benefit even unseen tasks (Pfeiffer et al., 2021; Huang et al., 2024a). Compared with whole-model merging, composing lightweight adapters requires more careful knowledge selection and combination. Zhao et al. (2025b) address this issue via structure-aware merging instead of simple weighted combination, improving low-resource adaptation by aligning the merging operator with how the adapter interacts with the base model. Miyano and Arase (2025) further combine parameter updating and pruning to reduce noise in merging. Beyond static composition, Cheng et al. (2025b) and Li et al. (2025a) extend adapter composition to dynamic settings by conditioning composition on the input and further refining it at the rank level of LoRA.

Takeaway: How can we perform effective model updating under limited labeled data?
§4.1: Maintaining **stable optimization** under scarce supervision by refining strategies of post-training $\phi(\cdot)$;
§4.2: Exploiting **transferable supervision** from the target task data $D_A^{1,t}$, data from other tasks $\{D_A^{2,t}, ..., D_A^{n,t}\}$, and transferable training signals from historical models $\{M^0, ..., M^{t-1}\}$ and auxiliary models $\mathbb{M}_R$;
§4.3: Enhancing adaptation through **model collaboration** across available checkpoints $\{M^0, \ldots, M^{t-1}\}$ and the additional models $\mathbb{M}_R$ at different neural granularities.

5 Reliable Inference

Reliable inference aims to improve the trustworthiness of model predictions under limited task-adaptation data. Existing studies improve the generation process with external resources, verification, and inference strategies. Although not always designed for low-resource settings, these works are highly relevant because they help reduce hallucinations and improve reliability. These efforts can be viewed as improving the generation function $\xi(\cdot)$.

5.1 External-Resource-Augmented Inference

External-resource-augmented inference improves generation by introducing additional support from external knowledge or tools, enabling the adapted model to access relevant evidence and problem-solving support beyond its parametric knowledge.

Retrieval-Augmented Generation. Existing work on RAG primarily focuses on improving retrievers on target tasks, determining trigger points of retrieval, and effectively integrating evidence into reasoning. Regarding retrievers, Lee et al. (2025), Siriwardhana et al. (2023) and Mao et al. (2024) find a major performance decline under domain shift and suggest introducing domain-specific signals for better retriever adaptation. For retrieval triggering, Huang et al. (2025a) and Yao et al. (2025) explore

how internal representations or uncertainty signals determine the necessity of retrieval. Wang et al. (2025e) and Zhang et al. (2025e) further introduce explicit gating and knowledge boundary modeling for fine-grained control over retrieval decisions, while Tang et al. (2025) employ a multi-armed bandit framework to adaptively select among retrieval strategies. After retrieval, effectively coupling evidence with reasoning is also crucial. Islam et al. (2024) focus on mitigating misleading distractors in retrieved results, whereas Tran et al. (2025a) further embed retrieval into multi-step reasoning.

Tool-Augmented Generation. Compared with conventional RAG, tool-augmented generation has shown stronger potential in underrepresented programming languages (Singha et al., 2025). However, Kulkarni et al. (2025) also point out the fragility of simple tool usage in some low-resource settings. Existing work on tool-augmented generation under scarce supervision mainly focuses on tool retrieval and the acquisition of tool-use abilities in LLMs. For the first line, Chen et al. (2024), He et al. (2025), and Sengupta et al. (2026) explore the feasibility of enhancing tool retrievers by additional supervised signals from synthetic queries, simulated tools, etc. For the second line, Schick et al. (2023) first demonstrate that LLMs can learn and acquire the abilities of API usage. In follow-up works, Luo et al. (2025c) and Zeng et al. (2025a) further explore how self-training and reinforcement learning teach LLMs to use tools without relying on abundant tool-use demonstrations.

5.2 Reasoning-Enhanced Inference

For low-resource adaptation, limited supervision often leaves the adapted model with unstable reasoning behaviors. Reasoning-enhanced inference addresses this via effective reasoning verification, or expanding the reasoning budget with more sampling and deeper deliberation for robust responses.

Verification-Based Inference. In the era of GPT-3.5 and GPT-4, Madaan et al. (2023) and Gero et al. (2023) demonstrate the possibility of improving model outputs with self-feedback and refinement. Zhang et al. (2024c) further explore how self-generated critique data can be used to fine-tune self-correcting reasoners. However, Huang et al. (2024b) and Hong et al. (2024) also highlight the fragility of relying solely on intrinsic self-verification. To address this issue, Gou et al. (2024) introduce tool-interactive critiquing, leveraging external tool feedback to alleviate distortions in self-evaluation. From another perspective, Kawabata and Sugawara (2024) and He et al. (2024) emphasize that verification should operate on the reasoning process rather than only the final answer. Zhu et al. (2025a) further address the out-of-distribution issue of process verification by introducing a retrieval-enhanced mechanism for reasoning evaluation.

Test-Time Scaling. Early work shows that aggregating multiple reasoning paths yields more reliable predictions (Wang et al., 2023b), inspiring later test-time scaling methods with more budget (Tran et al., 2025b; Wang et al., 2025b). However, recent work also shows that increasing reasoning computation is not always beneficial. Zhao et al. (2025c) emphasize that reasoning depth should be adapted to question complexity, while Zeng et al. (2025b) suggest that longer reasoning traces may even reflect overthinking, especially harmful for under-adapted models. Recent work also explores how to better combine or allocate scaling strategies: Chang et al. (2025) integrate sequential and parallel scaling with step-level verifier guidance, whereas Wang et al. (2025h) focus on optimizing resource allocation under latency constraints. In addition, Liu et al. (2025d) highlight the need to coordinate prompting strategies with scaling behavior.

5.3 Risk-Controlled Inference

Even with external resources and reasoning enhancement, models adapted with limited supervision may still produce unreliable output. Risk-controlled inference addresses this by allowing the model to respond selectively based on uncertainty.

Abstention and Deferral. Li et al. (2025b) argue that unclear knowledge boundaries are a key source of hallucination in LLMs, suggesting risk-control mechanisms such as abstention and deferral. Yang et al. (2023) address this issue during alignment by encouraging models to better recognize and respect their knowledge boundaries, whereas Feng et al. (2024a) explore a training-free inference strategy that improves abstention via multilingual feedback signals. Wen et al. (2024) and Machcha et al. (2026) further propose abstention-oriented evaluation frameworks. Beyond abstention, Farinhas et al. (2025) and Soiffer et al. (2025) adopt deferral methods by estimating output reliability and routing difficult queries to stronger models.

Selective Prediction. In risk-controlled inference, a key question is whether the model should answer instantly. Early work successfully approaches this

issue with confidence-aware evaluation (Yoshikawa and Okazaki, 2023; Chen et al., 2023b). Jurayj et al. (2025) further apply selective prediction to test-time scaling, mitigating unreliable behavior from deeper reasoning. Beyond raw confidence scores, Wang et al. (2025g) introduce conformal uncertainty, using calibration-based significance testing to provide more robust risk control. In addition, Vashurin et al. (2025b) propose a length-invariant post-hoc correction to relieve the length-based uncertainty bias, while Vashurin et al. (2025a) provide a unified benchmark for uncertainty quantification, which contributes not only to selective prediction but also to other tasks like hallucination detection.

> ***Takeaway: How can we improve reliability at inference time for low-resource LLMs?***
> §5.1: Augmenting inference with **additional evidence** from external knowledge or tools base $\mathbb{D}_K$;
> §5.2: Mitigating reasoning mistakes by **checking and refining** intermediate states of $\xi(\cdot)$, optionally with support from external resources $\mathbb{D}_K$ and effective reasoning strategies r;
> §5.3: Managing output risks via **flexible response control**, based on uncertainty estimated from the internal states of $\xi(\cdot)$ and diverse outputs induced by reasoning strategies r.

6 Challenges and Future Directions

This section discusses the challenges and future directions intrinsic to low-resource task adaptation, focusing on adaptation under limited supervision instead of investigating all related research lines.

Target-Oriented Knowledge Transfer. Existing studies on adaptive data mixing and multi-task adaptation have shown that auxiliary tasks can improve transfer (Li et al., 2025c; Zhang et al., 2025b), but low-resource task adaptation requires more than broad multi-task capability. In many practical scenarios, the goal is not to build a uniformly strong multi-task model, but to obtain a model that performs well on one or a few particular tasks with limited supervision. This raises a target-oriented transfer problem: how to identify which tasks or data should be emphasized, down-weighted, or even discarded during adaptation. Future work should move beyond static positive-transfer assumptions and develop dynamic strategies that refine the training objective towards the specific target tasks.

Capability Preconditions for Task Adaptation. Many low-resource adaptation methods implicitly assume that the adapted model already has a certain level of target-task competence. For example, semi-supervised expansion requires reliable pseudo-label generation (Luo et al., 2025a; Park and Caragea, 2024), and test-time scaling assumes that additional computation can improve reasoning (Chang et al., 2025; Zhao et al., 2025c). When the adapted model performs poorly on the target task, these methods may fail to provide gains or even amplify errors. Future research should characterize the capability thresholds under which such bootstrapping methods become effective, and investigate whether domain adaptation, auxiliary-task transfer, or external evidence can provide the minimum capability needed for target-task specialization.

Sufficient Evidence under Resource Scarcity. Retrieval- and tool-augmented generation provide a natural way to compensate for missing parametric knowledge (Tran et al., 2025a; Sengupta et al., 2026). However, their effectiveness also depends on the availability of external resources, such as task-specific instances, tool descriptions, and structured knowledge. For low-resource tasks, such resources may themselves be sparse and poorly aligned with the target task. A promising direction is to connect data-centric methods with inference-time augmentation, studying how synthetic data generation and data selection can be used to build and refine external evidence for reliable inference.

Comprehensive Benchmarks for Underexplored Low-Resource Tasks. Progress in low-resource task adaptation is constrained by the limited visibility of related tasks[2]. Many underrepresented tasks remain understudied, with datasets scattered across specialized communities rather than unified evaluation frameworks. Current evaluations often focus on simple sentiment analysis or general QA, while more challenging tasks, such as complex reasoning and agents, remain less explored or show limited gains. This motivates comprehensive benchmarks that cover diverse domains and task formats, specify resource profiles of both task-specific data and external resources, and evaluate robustness, and evidence faithfulness beyond accuracy.

7 Conclusions

We have presented a systematic survey of low-resource task adaptation in LLMs, spanning data-centric methods, effective post-training, and reliable inference. This taxonomy highlights how limited supervision shapes the full adaptation pipeline and exposes key challenges for reliable deployment under resource constraints. Beyond organizing existing work, this survey aims to spur further progress toward resource-aware NLP for underrepresented tasks and communities.

[2] Appendix A.2 and A.3 review low-resource tasks in depth.

Limitations

Boundary of the Survey Scope. This survey focuses on low-resource task adaptation as the final specialization stage before applying LLMs to downstream tasks. In particular, we emphasize the scarcity of in-domain, task-specific labeled data, which is often difficult to obtain but crucial for effective task adaptation. This focus allows us to examine a practical yet underexplored problem, but it also makes the survey scope relatively specific. Since works that directly target low-resource task adaptation remain limited, we also include supporting studies from related areas (e.g., instruction tuning and alignment), when they provide transferable insights for addressing scarce task-specific supervision. To make this distinction explicit, we mark relatively direct works in Figure 3 and provide a detailed discussion of the survey scope in Appendix A.1. In addition, this survey mainly focuses on text-based LLMs. Therefore, how multimodal information can be effectively exploited for task adaptation under scarce supervision remains an important topic for future surveys.

Lack of Unified Empirical Comparison. A unified empirical comparison across data-centric, effective post-training, and reliable inference methods would be highly valuable. However, existing studies are limited and mainly conducted under diverse target tasks, data scales, and foundation models, making direct comparison of reported results across papers unfair. Therefore, this survey is positioned primarily as a methodological guide instead of a benchmark-style empirical ranking. We aim to summarize what kinds of methods can be considered when developing LLMs for low-resource tasks. Nevertheless, we believe building unified benchmarks and conducting controlled comparisons of low-resource task adaptation remain critically important. Appendix A.2 shows the common low-resource tasks experimented with in current research, while Appendix A.3 further presents a series of promising tasks that deserve more attention in future low-resource task benchmarks.

Limited Use of Formalized Analysis. In the preliminary and taxonomy section (§2), we formalize low-resource task adaptation through the data-centric procedure $\chi(\cdot)$, post-training operator $\phi(\cdot)$, and inference process $\xi(\cdot)$, aiming to clarify the available inputs and methodological roles of different stages. These formulations are used as an organizing framework and are further revisited in the *Takeaways* for easier understanding of how different kinds of approaches address the low-resource issue. Nevertheless, due to the page limitation and the diversity of the surveyed methods, we cannot provide a detailed formalization for every individual work. Researchers interested in method-specific details are encouraged to consult the original papers or more specialized surveys on each subfield. We provide a method comparison of core methods in Appendix A.1, with representative surveys on specific methodological threads listed in Appendix A.5.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, and 26 others. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Trans. Mach. Learn. Res.*, 2024.

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. RAFT: A real-world few-shot text classification benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, William Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020.

TICO-19: the translation initiative for covid-19. *CoRR*, abs/2007.01788.

Trapit Bansal, Salaheddin Alzubi, Tong Wang, Jay-Yoon Lee, and Andrew McCallum. 2022. Meta-adapters: Parameter efficient few-shot fine-tuning through meta-learning. In *International Conference on Automated Machine Learning, AutoML 2022, 25-27 July 2022, Johns Hopkins University, Baltimore, MD, USA*, Proceedings of Machine Learning Research, pages 19/1–18. PMLR.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. TLDR: extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 4766–4777. Association for Computational Linguistics.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. 2022. A scalable and extensible approach to benchmarking nl2code for 18 programming languages. *CoRR*, abs/2208.08227.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

Kaiyan Chang, Yonghao Shi, Chenglong Wang, Hang Zhou, Chi Hu, Xiaoqian Liu, Yingfeng Luo, Yuan Ge, Tong Xiao, and JingBo Zhu. 2025. Step-level verifier-guided hybrid test-time scaling for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 18462–18477. Association for Computational Linguistics.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. 2025. Text-to-lora: Instant transformer adaption. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023a. An empirical survey of data augmentation for limited data learning in NLP. *Trans. Assoc. Comput. Linguistics*, 11:191–211.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Ö. 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, Singapore, December 6-10, 2023*, Findings of ACL, pages 5190–5213. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Xin Chen, Junchao Wu, Shu Yang, Runzhe Zhan, Zeyu Wu, Min Yang, Shujian Huang, Lidia S. Chao, and Derek F. Wong. 2026a. Neuron-aware data selection in instruction tuning for large language models. In *The Fourteenth International Conference on Learning Representations*.

Yanfei Chen, Jinsung Yoon, Devendra Singh Sachan, Qingze Wang, Vincent Cohen-Addad, Mohammad-Hossein Bateni, Chen-Yu Lee, and Tomas Pfister. 2024. Re-invoke: Tool invocation rewriting for zero-shot tool retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 4705–4726. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Findings of ACL, pages 5062–5074. Association for Computational Linguistics.

Zhuo Chen, Yuxuan Miao, and Deyi Xiong. 2026b. Data mixing for large language models pretraining: A survey and outlook. *Data Intell.*, 8(1):15–55.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025a.

Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 10400–10408. ijcai.org.

Ming Cheng, Jiaying Gong, and Hoda Eldardiry. 2025b. Sci-lora: Mixture of scientific loras for cross-domain lay paraphrasing. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 18524–18541. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 1679–1705. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mac. Intell.*, 5(3):220–235.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*. Asian Federation of Natural Language Processing.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Jennifer D'souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. SemEval-2025 task 5: LLMs4Subjects - LLM-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2570–2583, Vienna, Austria. Association for Computational Linguistics.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Ramaneswaran S., Utkarsh Tyagi, and Dinesh Manocha. 2024. Coda: Constrained generation based data augmentation for low-resource NLP. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, volume NAACL 2024 of *Findings of ACL*, pages 3754–3769. Association for Computational Linguistics.

Brett Hemenway Falk and Gerry Tsoukalas. 2026. The ai layoff trap. *Preprint*, arXiv:2603.20617.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on*

Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024, pages 6491–6501. ACM.

António Farinhas, Nuno Miguel Guerreiro, Sweta Agrawal, Ricardo Rei, and André F. T. Martins. 2025. Translate smart, not hard: Cascaded translation systems with quality-aware deferral. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 26742–26756. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4125–4150. Association for Computational Linguistics.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024b. Mixture-of-loras: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11371–11380. ELRA and ICCL.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, CEUR Workshop Proceedings. CEUR-WS.org.

Markus Frohmann, Carolin Holtermann, Shahed Masoudian, Anne Lauscher, and Navid Rekabsaz. 2024. Scalearn: Simple and highly parameter-efficient task transfer by learning to scale. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 11743–11776. Association for Computational Linguistics.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. 2025. Training-free LLM merging for multi-task learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 33111–33124. Association for Computational Linguistics.

Boyan Gao, Xin Wang, Yibo Yang, and David A. Clifton. 2025. Optimization-inspired few-shot adaptation for large language models. *CoRR*, abs/2505.19107.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Luis Gasco, Hermenegildo Fabregat, Laura García-Sardiña, Daniel Deniz, Alvaro Rodrigo, Paula Estrella, and Rabih Zbib. 2025. Talentclef at clef2025: Skill and job title intelligence for human capital management. In *Advances in Information Retrieval*, pages 479–486, Cham. Springer Nature Switzerland.

Luis Gasco Sanchez, Hermenegildo Fabregat, Laura Garcia-Sardina, and Rabih Zbib. 2026. Talentclef at clef2026: Skill and job title intelligence for human capital management. In *Advances in Information Retrieval*, pages 251–258, Cham. Springer Nature Switzerland.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xu Guo, Zilin Du, Boyang Li, and Chunyan Miao. 2024. Generating synthetic datasets for few-shot prompt tuning. In *First Conference on Language Modeling*.

Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu, and Sachindra Joshi. 2025. Selective self-to-supervised fine-tuning for generalization in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 6240–6249. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, 2024.

Jie He, Jennifer Neville, Mengting Wan, Longqi Yang, Hui Liu, Xiaofeng Xu, Xia Song, Jeff Z. Pan, and Pei Zhou. 2025. Gentool: Enhancing tool generalization in language models through zero-to-one and weak-to-strong simulation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 1097–1122. Association for Computational Linguistics.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. 2024. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2086–2099. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2208–2222. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 900–925. Association for Computational Linguistics.

S. K. Hong and Tae Young Jang. 2022. AMAL: meta knowledge-driven few-shot adapter learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10381–10389. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chengkai Huang, Yu Xia, Rui Wang, Kaige Xie, Tong Yu, Julian J. McAuley, and Lina Yao. 2025a. Embedding-informed adaptive retrieval-augmented generation of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 1403–1412. Association for Computational Linguistics.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024a. Lorahub: Efficient cross-task generalization via dynamic loRA composition. In *First Conference on Language Modeling*.

Hao Huang, Jingtao Ding, Mengqi Liao, Xin Wang, Jinyang Ban, Yuan Yuan, Huaiyu Wan, and Yong Li. 2026. LoRAGen: Structure-aware weight space learning for loRA generation. In *The Fourteenth International Conference on Learning Representations*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024b. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wei Huang, Anda Cheng, and Yinggui Wang. 2025b. Mitigating catastrophic forgetting in large language models with forgetting-aware pruning. In *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 21842–21856. Association for Computational Linguistics.

Yizheng Huang and Jimmy Xiangji Huang. 2026. A survey on retrieval-augmented text generation for large language models. *ACM Comput. Surv.* Just Accepted.

Yukun Huang, Kun Qian, and Zhou Yu. 2023. Learning a better initialization for soft prompts via meta-learning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 - Volume 2: Short Papers, Nusa Dua, Bali, November 1-4, 2023*, pages 67–75. Association for Computational Linguistics.

Tingfeng Hui, Zhenyu Zhang, Shuohuan Wang, Weiran Xu, Yu Sun, and Hua Wu. 2025. HFT: half fine-tuning for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12791–12819. Association for Computational Linguistics.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2025. Scaling laws for downstream task performance in machine translation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Shayekh Bin Islam, Md. Asib Rahman, K. S. M. Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md. Rizwan Parvez. 2024. Open-rag: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 14231–14244. Association for Computational Linguistics.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1393–1409. Association for Computational Linguistics.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Xinyi Dai, Yan Xu, Weinan Gan, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and Wei Wang. 2025. Instruction-tuning data synthesis from scratch via web reconstruction. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 6603–6618. Association for Computational Linguistics.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. 2025. Is that your final answer? test-time scaling improves selective question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 636–644. Association for Computational Linguistics.

Akira Kawabata and Saku Sugawara. 2024. Rationale-aware answer verification by pairwise self-evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16178–16196. Association for Computational Linguistics.

Soohyeong Kim, Seok Jun Hwang, JungHyoun Kim, Jeonghyeon Park, and Yong Suk Choi. 2025. Regraphrag: Reorganizing fragmented knowledge graphs for multi-perspective retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 5426–5443. Association for Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Mayank Kulkarni, Vittorio Mazzia, Judith Gaspers, Chris Hench, and Jack FitzGerald. 2025. Massive-agents: A benchmark for multilingual function-calling in 52 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20193–20215. Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Changho Lee, Janghoon Han, Seonghyeon Ye, Stanley Jungkyu Choi, Honglak Lee, and Kyunghoon Bae. 2024. Instruction matters: A simple yet effective task selection for optimized instruction tuning of specific tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18620–18642. Association for Computational Linguistics.

Dohyeon Lee, Jongyoon Kim, Jihyuk Kim, Seung-won Hwang, and Joonsuk Park. 2025. trag: Term-level retrieval-augmented generation for domain-adaptive retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long*

Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025, pages 6566–6578. Association for Computational Linguistics.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 666–684. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Dongyue Li, Ziniu Zhang, Lu Wang, and Hongyang Zhang. 2024a. Scalable fine-tuning from multiple data sources: A first-order approximation approach. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 5608–5623. Association for Computational Linguistics.

Guanyu Li, Zhiheng Xi, Zhihao Zhang, Boyang Hong, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025a. Loracoe: Improving large language model via composition-based lora expert. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 31290–31304. Association for Computational Linguistics.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14255–14273. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024c. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7602–7635. Association for Computational Linguistics.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025b. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5131–5157. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Xinzhe Li. 2025. A survey on LLM test-time compute via search: Tasks, LLM profiling, search algorithms, and relevant frameworks. *Trans. Mach. Learn. Res.*, 2025.

Yangning Li, Tingwei Lu, Yinghui Li, Yankai Chen, Wei-Chieh Huang, Wenhao Jiang, Hui Wang, Hai-Tao Zheng, and Philip S. Yu. 2025c. Teaching according to talents! instruction tuning llms with competence-aware curriculum learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 11724–11741. Association for Computational Linguistics.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025d. A survey of RAG-reasoning systems in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12120–12145, Suzhou, China. Association for Computational Linguistics.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024d. Empowering large language models for textual data augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 12734–12751. Association for Computational Linguistics.

Zhuang Li, Yuncheng Hua, Thuy-Trang Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. 2025e. SCAR: data selection via style consistency-aware response ranking for efficient instruction-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12756–12790. Association for Computational Linguistics.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *CoRR*, abs/2303.15647.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning RAG via system 1

or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pages 1954–1966. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Vijay Lingam, Atula Neerkaje, Aditya Vavre, Aneesh Shetty, Gautham Krishna Gudur, Joydeep Ghosh, Eunsol Choi, Alex Dimakis, Aleksandar Bojchevski, and Sujay Sanghavi. 2024. SVFT: parameter-efficient fine-tuning with singular vectors. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jinda Liu, Yi Chang, and Yuan Wu. 2025a. R-lora: Randomized multi-head lora for efficient multi-task learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 660–674. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 61–68. Association for Computational Linguistics.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025b. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 6107–6117. ACM.

Xiyang Liu, Chunming Hu, Richong Zhang, Junfan Chen, and Baowen Xu. 2025c. Improving data annotation for low-resource relation extraction with logical rule-augmented collaborative language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1497–1510. Association for Computational Linguistics.

Yajiao Liu, Congliang Chen, Junchi Yang, and Ruoyu Sun. 2026. Rethinking data mixture for large language models: A comprehensive survey and new perspectives. In *Findings of the Association for Computational Linguistics: EACL 2026, Rabat, Morocco, March 24-29, 2026*, Findings of ACL, pages 275–289. Association for Computational Linguistics.

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. 2025d. Rethinking the role of prompting strategies in LLM test-time scaling: A perspective of probability theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 27962–27994. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. *CoRR*, abs/2407.06089.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025a. Semi-supervised fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 2795–2808. Association for Computational Linguistics.

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and Ming Zhang. 2025b. A survey on efficient large language model training: From data-centric perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30904–30920. Association for Computational Linguistics.

Ne Luo, Aryo Pradipta Gema, Xuanli He, Emile van Krieken, Pietro Lesci, and Pasquale Minervini. 2025c. Self-training large language models for tool-use without demonstrations. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pages 1253–1271. Association for Computational Linguistics.

Chuancheng Lv, Lei Li, Shitou Zhang, Gang Chen, Fanchao Qi, Ningyu Zhang, and Hai-Tao Zheng. 2024. Hyperlora: Efficient cross-task generalization via constrained low-rank adapters generation. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, volume EMNLP 2024 of *Findings of ACL*, pages 16376–16393. Association for Computational Linguistics.

Qingsong Lv, Yangning Li, Zihua Lan, Zishan Xu, Jiwei Tang, Tingwei Lu, Yinghui Li, Wenhao Jiang, Hong-Gee Kim, Hai-Tao Zheng, and Philip S. Yu. 2025. RAISE: reinforced adaptive instruction selection for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 11708–11723. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. STAR: boosting low-resource information extraction by structure-to-text data generation with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18751–18759. AAAI Press.

Sravanthi Machcha, Sushrita Yerra, Sahil Gupta, Aishwarya Sahoo, Sharmin Sultana, Hong Yu, and Zonghai Yao. 2026. Knowing when to abstain: Medical llms under clinical uncertainty. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 6153–6182. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *CoRR*, abs/2212.10511.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. Rag-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 725–735. Association for Computational Linguistics.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 216–223. European Language Resources Association (ELRA).

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 10456–10470. Association for Computational Linguistics.

Paramita Mirza, Lucas Weber, and Fabian Küch. 2025. Stratified selective sampling for instruction tuning with dedicated scoring strategy. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 19949–19974. Association for Computational Linguistics.

Ryota Miyano and Yuki Arase. 2025. Adaptive lora merge with parameter pruning for low-resource generation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 19353–19366. Association for Computational Linguistics.

Hyeonseok Moon, Jaehyung Seo, Seonmin Koo, Jinsung Kim, Young-kyoung Ham, Jiwon Moon, and Heuiseok Lim. 2025. Limacost: Data valuation for instruction tuning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 12841–12854. Association for Computational Linguistics.

Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 12585–12611. Association for Computational Linguistics.

Thao Nguyen, Jeffrey Li, Sewoong Oh, Ludwig Schmidt, Jason Weston, Luke Zettlemoyer, and Xian Li. 2024. Better alignment with instruction back-and-forth translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 13289–13308. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024. SemEval task 1: Semantic textual relatedness for African and Asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1963–1978, Mexico City, Mexico. Association for Computational Linguistics.

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 F 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. Curran Associates, Inc.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: *Surveying the Landscape of Diverse Automated Correction Strategies*. *Trans. Assoc. Comput. Linguistics*, 12:484–506.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Tanmay Parekh, Yuxuan Dong, Lucas Bandarkar, Artin Kim, I-Hung Hsu, Kai-Wei Chang, and Nanyun Peng. 2025. Snare: Domain-aware data generation for low-resource event detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20583–20604. Association for Computational Linguistics.

Seoyeon Park and Cornelia Caragea. 2024. Verifymatch: A semi-supervised learning paradigm for natural language inference with confidence-aware mixup. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19319–19335. Association for Computational Linguistics.

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing low-resource llms classification with PEFT and synthetic data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 6017–6023. ELRA and ICCL.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podrouzek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2498–2511, Vienna, Austria. Association for Computational Linguistics.

Carla Perez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 487–503. Association for Computational Linguistics.

Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. 2023. Hypertuning: Toward adapting large language models without back-propagation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 27854–27875. PMLR.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alipio Mario Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Guimaraes, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval 2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2610–2643, Vienna, Austria. Association for Computational Linguistics.

Shanbao Qiao, Xuebing Liu, Akshat Gupta, and Seung-Hoon Na. 2025. Seqmmr: Sequential model merging and LLM routing for enhanced batched sequential

knowledge editing. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 16932–16947. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025a. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi R. Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, and 24 others. 2025b. Tool learning with foundation models. *ACM Comput. Surv.*, 57(4):101:1–101:40.

Benedictus Kent Rachmat, Thomas Gerald, Zheng Zhang, and Cyril Grouin. 2025. QA analysis in medical and legal domains: A survey of data augmentation in low-resource settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1132–1144. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3066–3079. Association for Computational Linguistics.

Jun Rao, Xuebo Liu, Lian Lian, Shengjun Cheng, Yunjie Liao, and Min Zhang. 2024. Commonit: Commonality-aware instruction tuning for large language models via data partitions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10064–10083. Association for Computational Linguistics.

H. S. V. N. S. Kowndinya Renduchintala, Sumit Bhatia, and Ganesh Ramakrishnan. 2024. SMART: submodular data mixture strategy for instruction tuning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 12916–12934. Association for Computational Linguistics.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Gözde Gül Sahin. 2022. To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP. *Comput. Linguistics*, 48(1):5–42.

Ixak Sarasua, Ander Corral, and Xabier Saralegi. 2025. Diploma: Efficient adaptation of instructed llms to low-resource languages via post-training delta merging. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 24898–24912. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3687–3697. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Fabian David Schmidt, Philipp Borchert, Ivan Vulic, and Goran Glavas. 2024a. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 6724–6743. Association for Computational Linguistics.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024b. Prompting-based synthetic data generation for few-shot question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13168–13178. ELRA and ICCL.

Saptarshi Sengupta, Zhengyu Zhou, Jun Araki, Xingbo Wang, Bingqing Wang, Suhang Wang, and Zhe Feng. 2026. Tooldreamer: Instilling LLM reasoning into tool retrievers. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 5465–5482. Association for Computational Linguistics.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2026. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3):63:1–63:38.

Reece S Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2025. LoRA vs full fine-tuning: An illusion of equivalence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ozge Mercanoglu Sincan, Júlio C. S. Jacques Júnior, Sergio Escalera, and Hacer Yalim Keles. 2021. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3472–3481. Computer Vision Foundation / IEEE.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Ananya Singha, Mukul Singh, Hosein Hasanbeig, Arjun Radhakrishna, and Sumit Gulwani. 2025. Adapting language models for low-resource programming languages. In *NeurIPS 2025 Fourth Workshop on Deep Learning for Code*.

Shamane Siriwardhana, Rivindu Weerasekera, Tharindu Kaluarachchi, Elliott Wen, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Trans. Assoc. Comput. Linguistics*, 11:1–17.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Duncan Soiffer, Steven Kolawole, and Virginia Smith. 2025. Semantic agreement enables efficient open-ended LLM cascades. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 2499–2537. Association for Computational Linguistics.

Patrick Sutanto, Joan Santoso, Esther Irawati Setiawan, and Aji Prasetya Wibawa. 2025. LLM distillation for efficient few-shot multiple choice question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 8502–8530. Association for Computational Linguistics.

Marton Szep, Daniel Rueckert, Rüdiger von Eisenhart-Rothe, and Florian Hinterwimmer. 2026. Fine-tuning large language models with limited data: A survey and practical guide. *Trans. Assoc. Comput. Linguistics*, 14:341–377.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 930–957. Association for Computational Linguistics.

Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and Sihong Xie. 2025. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 3248–3254. Association for Computational Linguistics.

Mingxu Tao, Chen Zhang, Quzhe Huang, Tianyao Ma, Songfang Huang, Dongyan Zhao, and Yansong Feng. 2024. Unlocking the potential of model merging for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 8705–8720. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. Semeval-2026 task 6: CLARITY - unmasking political question evasions. *CoRR*, abs/2603.14027.

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. *Preprint*, arXiv:2302.13971.

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025a. RARE: retrieval-augmented reasoning enhancement for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1,*

2025, pages 18305–18330. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025b. Scaling test-time compute for low-resource languages: Multilingual reasoning in llms. *CoRR*, abs/2504.02890.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025a. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Trans. Assoc. Comput. Linguistics*, 13:220–248.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, and Maxim Panov. 2025b. UNCERTAINTY-LINE: length-invariant estimation of uncertainty for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 7881–7908. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 13697–13720. Association for Computational Linguistics.

Bowen Wang, Haiyuan Wan, Liwen Shi, Chen Yang, Peng He, Yue Ma, Haochen Han, Wenhao Li, Tiao Tan, Yongjian Li, Fangming Liu, Yifan Gong, and Sheng Zhang. 2025a. RECALL: representation-aligned catastrophic-forgetting alleviation via hierarchical model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 16381–16395. Association for Computational Linguistics.

Jianing Wang, Qiushi Sun, Nuo Chen, Chengyu Wang, Jun Huang, Ming Gao, and Xiang Li. 2023a. Uncertainty-aware parameter-efficient self-training for semi-supervised language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 7873–7884. Association for Computational Linguistics.

Kaixin Wang, Tianlin Li, Xiaoyu Zhang, Aishan Liu, Xianglong Liu, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Bin Shi. 2025b. Codechemist: Functional knowledge transfer for low-resource code generation via test-time scaling. *CoRR*, abs/2510.00501.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025c. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artif. Intell. Rev.*, 58(8):227.

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. 2025d. Data whisperer: Efficient data selection for task-specific LLM fine-tuning via few-shot in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23287–23305. Association for Computational Linguistics.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025e. Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 491–503. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Awadallah, and Jianfeng Gao. 2022a. List: Lite prompted self-training makes parameter-efficient few-shot learners. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, volume NAACL 2022 of *Findings of ACL*, pages 2262–2281. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max

Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yueqian Wang, Chang Liu, Kai Chen, Xi Wang, and Dongyan Zhao. 2022b. SMASH: improving small language models' few-shot ability with prompt-based distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Findings of ACL, pages 6608–6619. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi N. Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024c. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19519–19529. Association for Computational Linguistics.

Zhenyi Wang, Yapeng Jia, Haiyan Ning, Peng Wang, Dan Wang, and Yitao Cao. 2025f. HMCL: task-optimal text representation adaptation through hierarchical contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 13495–13518. Association for Computational Linguistics.

Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. 2024d. What are tools anyway? a survey from the language model perspective. In *First Conference on Language Modeling*.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025g. Sconu: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19052–19075. Association for Computational Linguistics.

Zili Wang, Tianyu Zhang, Haoli Bai, Lu Hou, Xianzhi Yu, Wulong Liu, Shiming Xiang, and Lei Zhu. 2025h. Faster and better llms via latency-aware test-time scaling. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 17124–17137. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jingyu Wei, Bo Liu, Tianjiao Wan, Baoyun Peng, Xingkong Ma, and Mengmeng Guo. 2025. Ji^{2}s: Joint influence-aware instruction data selection for efficient fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 514–527. Association for Computational Linguistics.

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, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 3437–3450. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Trans. Assoc. Comput. Linguistics*, 13:529–556.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, and Reza Haf. 2024. Mixture-of-skills: Learning to optimize data usage for fine-tuning large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14226–14240, Miami, Florida, USA. Association for Computational Linguistics.

Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. 2025. ROSE: A reward-oriented data selection framework for LLM task-specific instruction

tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 13200–13219. Association for Computational Linguistics.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lingling XU, Haoran XIE, {S. Joe} QIN, Xiaohui TAO, and {Fu Lee} WANG. 2026. Parameter-efficient fine-tuning methods for pretrained language models : A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Publisher Copyright: © 1979-2012 IEEE.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. 2025. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *Transactions on Machine Learning Research*. Survey Certification.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025b. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Trans. Intell. Syst. Technol.*, 16(6):143:1–143:27.

Enneng Yang, Li Shen, Guibing Guo, Xing-Wei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2026. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Comput. Surv.*, 58(8):216:1–216:41.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *CoRR*, abs/2312.07000.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1028–1043. Association for Computational Linguistics.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2025. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 27022–27043. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7163–7189. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

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, Dubrovnik, Croatia, May 2-6, 2023*, volume EACL 2023 of *Findings of ACL*, pages 1972–1983. Association for Computational Linguistics.

Yirong Zeng, Xiao Ding, Yutai Hou, Yuxian Wang, Li Du, Juyi Dai, Qiuyang Ding, Duyu Tang, Dandan Tu, Weiwen Liu, Bing Qin, and Ting Liu. 2025a. Tool zero: Training tool-augmented llms via pure RL from scratch. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 9135–9147. Association for Computational Linguistics.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025b. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4651–4665. Association for Computational Linguistics.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024a. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. 2025a. A survey on data selection for LLM instruction tuning. *J. Artif. Intell. Res.*, 83.

Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. 2025b. More: A mixture of low-rank

experts for adaptive multi-task learning. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 1311–1324. Association for Computational Linguistics.

Qi Zhang, Yiming Zhang, Haobo Wang, and Junbo Zhao. 2024b. RECOST: external knowledge guided data-efficient instruction tuning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 10911–10921. Association for Computational Linguistics.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025c. A survey on test-time scaling in large language models: What, how, where, and how well? *Preprint*, arXiv:2503.24235.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025d. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 3080–3118. Association for Computational Linguistics.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024c. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 15637–15653. Association for Computational Linguistics.

Zhen Zhang, Xinyu Wang, Yong Jiang, Zile Qiao, Zhuo Chen, Guangyu Li, Feiteng Mu, Mengting Hu, Pengjun Xie, and Fei Huang. 2025e. KBM: delineating knowledge boundary for adaptive retrieval in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 21771–21782. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023b. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 943–956. Association for Computational Linguistics.

Chenyang Zhao, Xueying Jia, Vijay Viswanathan, Graham Neubig, and Tongshuang Wu. 2024. Self-guide: Better task-specific instruction following via self-synthetic finetuning. In *First Conference on Language Modeling*.

Shu Zhao, Shiji Yang, Shicheng Tan, Zhen Yang, Congyao Mei, Zhen Duan, Yanping Zhang, and Jie Chen. 2025a. Prompt contrastive transformation: An enhanced strategy for efficient prompt transfer in natural language processing. *Trans. Assoc. Comput. Linguistics*, 13:848–860.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025b. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9785–9800. Association for Computational Linguistics.

Zhengyi Zhao, Shubo Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. 2025c. T^2: An adaptive test-time scaling strategy for contextual question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 3731–3756. Association for Computational Linguistics.

Kai Zheng, Qingfeng Sun, Yaming Yang, Tengchao Lv, Yeyong Pi, Changlin Zhao, Fei Xu, and Qi Zhang. 2023. Adversarial knowledge stimulated contrastive prompting for few-shot language learners. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 13495–13507. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nat.*, 634(8032):61–68.

Jiachen Zhu, Congmin Zheng, Jianghao Lin, Kounianhua Du, Ying Wen, Yong Yu, Jun Wang, and Weinan Zhang. 2025a. Retrieval-augmented process reward model for generalizable mathematical reasoning. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August*

1, 2025, volume ACL 2025 of *Findings of ACL*, pages 8453–8468. Association for Computational Linguistics.

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024a. Multilingual large language models: A systematic survey. *CoRR*, abs/2411.11072.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025b. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 8912–8924. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024b. A survey on model compression for large language models. *Trans. Assoc. Comput. Linguistics*, 12:1556–1577.

A Appendix

The appendix is organized as follows:

- Appendix A.1 defines the scope of the research covered in this survey, compares representative works in terms of methodological design and experimental settings, and clarifies the boundaries between related categories in our taxonomy.
- Appendix A.2 covers common low-resource tasks and datasets in current research.
- Appendix A.3 discusses community perspectives on the needs of low-resource tasks and also provides an overview of available data and evaluation resources.
- Appendix A.4 reviews research communities and workshops related to low-resource NLP.
- Appendix A.5 complements this survey by summarizing representative surveys on independent methodological threads related to low-resource task adaptation.

Moreover, we maintain a continuously updated bibliography[3] and tutorial-style slides/notes[4] on low-resource adaptation in LLMs, intended to support ongoing and future research in this area.

A.1 Scope of Included Studies

The recent rise of generative LLMs has shifted much of the community's attention toward general-purpose instruction-following models, alignment, and prompting. As shown in Figure 4, the proportion of low-resource papers has significantly decreased after 2022. This observation does not imply that low-resource problems have been solved. Rather, Szep et al. (2026) and Zhou et al. (2024) observe that scaling model size, pretraining corpora, and computational resources cannot easily remove the need for effective adaptation under scarce supervision. Data scarcity remains a practical challenge for LLM deployment in low-resource languages, domains, and task-specific requirements.

[3]This list of papers adopts relatively relaxed criteria for topic coverage and is intended to serve as an inclusive reference for low-resource task adaptation. Please find the anonymous version of the online bibliography: https://anonymous.4open.science/r/A-Survey-of-Low-Resource-Task-Adaptation-in-LLMs. This bibliography is organized as a research resource under CC BY 4.0, and does not redistribute related papers. Please cite our survey when using this list.

[4]The slides/notes will also be released after publication to follow the anonymized review rules.

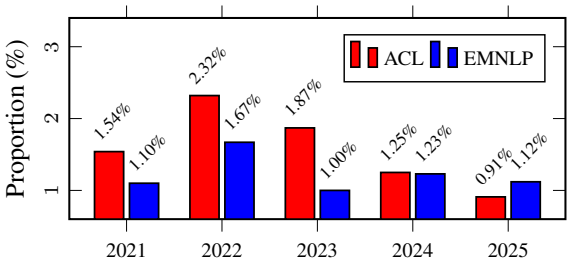

Figure 4: The proportion of papers related to low-resource tasks published at ACL and EMNLP 2021–2025. The detailed paper list can be found here.

Hence, this survey focuses on task adaptation for text-based LLMs under limited task-specific supervision (see Figure 6). We distinguish task adaptation from domain adaptation. Domain adaptation primarily exploits domain-relevant unlabeled data to improve domain familiarity, whereas task adaptation uses task-specific labeled data to optimize the model toward a target prediction objective. We choose task adaptation as the central scope because such labeled data are often among the most costly, scarce, and deployment-critical resources. Addressing low-resource challenges at the task-adaptation stage is therefore essential for improving the deployability of LLMs on low-resource tasks.

Based on this scope, we organize existing studies on low-resource task adaptation into a stage-based taxonomy (see Figure 3), and thoroughly compare representative works in Table 1, Table 2, and Table 3. Note that some methods combine multiple strategies to mitigate scarce supervision, and the boundaries between categories are not always mutually exclusive. We therefore clarify overlaps here to make our categorization criteria more explicit:

Adaptive Mixing vs. Multi-Task Learning. Data mixing uses data from multiple sources to broaden and rebalance the supervision signals available for low-resource task adaptation. Since these sources may include different tasks, data mixing is naturally related to multi-task learning, especially in adaptive mixing, where the proportions of tasks, datasets or instances are dynamically adjusted during training. In this survey, we distinguish them by their primary methodological focus: adaptive mixing is organized from a data-centric perspective, emphasizing the granularity and scheduling of mixture construction (Wu et al., 2024; Lv et al., 2025; Li et al., 2025c), whereas multi-task learning is discussed under effective post-training, focusing on how training across multiple tasks enables

Table 1: Overview of representative data-centric studies.

Work	Type	Method Input			Experimental Setting			Link
		Data Source	Auxiliary Model	Unlabeled Data	Task	Dataset	Foundation Model	
STAR (Ma et al., 2024)	Data Augmentation	Single	✓	×	Event Extraction; Relation Extraction	ACE05; TACRED	GPT-3.5; GPT-4	N/A
CoDa (Evuru et al., 2024)	Data Augmentation	Single	✓	×	Text Classification; NER; QA	HuffPost; Yahoo; OTS; ATIS; Massive; CoNLL-2003; OntoNotes-5.0; EBMNLP; BC2GM; SQuAD; NewsQA	LLaMA-13B; Mistral-7B; LLaMA-7B	link
Prompting-Based Synthetic Data Generation (Schmidt et al., 2024b)	Data Augmentation	Single	✓	× (Unlabeled data serves as contexts instead of direct annotation targets)	QA	Few-Shot MRQA Benchmark	T5 v1.1-large	link
DawGen (Guo et al., 2024)	Data Augmentation	Single	✓	×	Text Classification; Semantic Understanding	QQP; MRPC; MNLI; SNLI; QNLI; RTE; SICK	CTRL-1.6B; T5-large; Flan-T5-large	link
SEMIEVOL (Luo et al., 2025a)	Data Augmentation	Single	✓	✓	QA; Reasoning; Sentiment Analysis	MMLU; MMLU-Pro; ARC; FPB; USMLE; PubMedQA; ConvFinQA	GPT-4o-mini; LLaMA-3-8B	link
VerifyMatch (Park and Caragea, 2024)	Data Augmentation	Single	✓	✓	Text Classification	RTE; SICK; SNLI-2.5k; MNLI-2.5k	LLaMA-3-8B-Instruct; BERT-base	N/A
SNARE (Parekh et al., 2025)	Data Quality Evaluation	Single	× (Applied models are not for quality evaluation)	✓	Event Detection	ACE; SPEED; GENIA	LLaMA-3-8B-Instruct; LLaMA-3-70B-Instruct; GPT-3.5	link
Data Whisperer (Wang et al., 2025d)	Data Quality Evaluation	Single	✓	×	QA; Summarization; Reasoning	BioInstruct; DialogSum; GSM8K	LLaMA-3-8B-Instruct; Qwen-2.5-7B-Instruct; Mistral-Nemo-Instruct-2407; Qwen-2.5-3B-Instruct; Mistral-7B-Instruct-v0.2	link
ROSE (Wu et al., 2025)	Data Quality Evaluation	Single	✓	×	Preference Alignment	SHP; Stack Exchange; HH-RLHF	LLaMA-2-7B; LLaMA-2-13B; LLaMA-3.1-8B; LLaMA-3.1-8B-Instruct	N/A
CLMA (Liu et al., 2025c)	Data Quality Evaluation	Single	✓	✓	Relation Extraction	SemEval-2010; TACRED; TACREV; Re-RACRED	GPT-3.5; Flan-T5-large	link
INSTA (Lee et al., 2024)	Data Mixing	Multiple	×	×	General Benchmarks; Text Classification	P3; NIV2; BIG-Bench; BIG-Bench Hard	T5-3B	link
MOSPEC (Wu et al., 2024)	Data Mixing	Multiple	✓	×	QA; Reasoning; Dialogue	MMLU; MT-Bench; GSM8K; MATH; M-Math	Qwen-1.5-0.5B; Gemma-2B; LLaMA-3-8B	N/A

knowledge transfer to the target task (Frohmann et al., 2024; Miyano and Arase, 2025).

Contrastive Learning vs. Data Augmentation vs. Multi-Task Learning. Contrastive learning mitigates supervision scarcity by constructing auxiliary contrastive signals. Since such signals may be built from generated, perturbed, or task-transformed examples (Zheng et al., 2023), contrastive learning can overlap with data augmentation. Some methods also introduce auxiliary objectives during parameter optimization (Mao and Yu, 2024), making them related to multi-task learning. In this survey, when data augmentation or multi-task learning mainly serves as a supporting mechanism for constructing contrastive supervision, we place the works under contrastive learning.

Verification-Based Inference vs. Tool-Augmented Generation. Verification-based inference improves reliability by checking intermediate reasoning steps, candidate answers, or final outputs. Since self-verification based only on a model's internal signals can be unreliable (Huang et al., 2024b; Hong et al., 2024), some methods introduce external tools to assess reasoning processes or generated answers (Gou et al., 2024). This makes them conceptually similar to tool-augmented generation (Singha et al., 2025; Kulkarni et al., 2025; Schick et al., 2023), as both may involve external tool use. For a clearer boundary, we clarify the overlap based on the goals of tool application. Tool-augmented generation mainly uses tools to acquire external information and execute operations, while verification-based inference uses tools to validate reasoning traces and candidate outputs.

Risk-Controlled Inference vs. Verification-Based Inference. Risk-controlled inference reduces deployment risks by selectively abstaining or deferring when the model is uncertain (Yoshikawa and Okazaki, 2023; Feng et al., 2024a; Wang et al., 2025g). It is related to verification-based inference because both may use reliability signals to assess reasoning processes and candidate outputs. However, verification-based inference mainly focuses on how to validate, while risk-controlled in-

Table 2: Overview of representative effective post-training studies.

Work	Type	Method Input		Experimental Setting			Link
		Data Source	Model / Module	Task	Dataset	Foundation Model	
OFA (Gao et al., 2025)	Parameter-Efficient Adaptation	Single	Single	Text Classification	SST-2; Emoc; TREC; AGNews; SST-5; MR; DBPedia; Subj; HateSp18	LLaMA-2-7B; LLaMA-3-8B; LLaMA-3-8B-Instruct; GPT-2-XL	N/A
$(IA)^3$ (Liu et al., 2022a)	Parameter-Efficient Adaptation	Single	Single	Text Generation; Classification	Held-Out T0 Tasks (COPA; H-SWAG; Story Cloze; ANLI; CB; RTE; WSC; Winogrande; WiC); RAFT Benchmark	T0-3B; T0-11B	link
SCALEARN (Frohmann et al., 2024)	Transfer-Enhanced Training	Multiple	Multiple (Parameter Modules from Previous Training Steps)	Text Classification; Semantic Understanding; QA	GLUE; SuperGLUE; HumSet	RoBERTa-base; RoBERTa-large; XLM-R-base; XLM-R-large	link
HyperTuning (Phang et al., 2023)	Transfer-Enhanced Training	Multiple	Multiple (Hypernetwork)	Text Classification; Semantic Understanding; Reasoning; General Benchmarks	P3; MetaICL; S-NI	T5-large; T5-XL	N/A
HyperLoRA (Lv et al., 2024)	Transfer-Enhanced Training	Multiple	Multiple (Hypernetwork)	Text Classification; Semantic Understanding; QA; General Benchmarks	P3; S-NI; SuperGLUE; BIG-Bench Hard	T5-large; T5-XL; LLaMA-2-7B	N/A
MetaPT (Huang et al., 2023)	Transfer-Enhanced Training	Multiple	Single	Sentiment Analysis	SST-5; SST-2; Amazon-5; Amazon-2; SemEval-2016	T5-base; T5-large	N/A
SMASH (Wang et al., 2022b)	Transfer-Enhanced Training	Multiple	Multiple (Teacher and Student Models)	Text Classification; Semantic Understanding	GLUE; SNLI; SST-5; MR; CR	RoBERTa-large; DistilRoBERTa-base	link
SELF-GUIDE (Zhao et al., 2024)	Transfer-Enhanced Training	Single	Multiple (Models from previous training steps)	Text Classification; Generation	Super-NaturalInstructions V2 Benchmark	Vicuna-7B-1.5	link
AlignInstruct (Mao and Yu, 2024)	Transfer-Enhanced Training	Single	Single	Machine Translation	OPUS-100; Flores-200	BLOOMZ-7B1; BLOOMZ-3B; BLOOMZ-1B1	N/A
AKSCP (Zheng et al., 2023)	Transfer-Enhanced Training	Single	Single	Text Classification; Semantic Understanding	SST-2; SST-5; Subj; CoLA ;REC; MNLI; SNLI; QNLI; QQP RT; MRP; MR; MPQ; CR	RoBERTa-base	N/A
Model Merging for Low-Resource Languages (Tao et al., 2024)	Composition Adaptation	Multiple	Multiple (Models Optimized for Different Objectives)	Text Classification, Semantic Understanding, and Reasoning	MGSM; SIB-200; Belebele; MLiC-Eval	LLaMA-2-7B	N/A
Adaptive LoRA Merge with Pruning (Miyano and Arase, 2025)	Composition Adaptation	Multiple	Multiple (Parameter Modules Optimized for Different Objectives)	Summarization	MIMIC-III; SciTLDR; Bloomberg; NLP/Medical Paper	LLaMA-3-8B-Instruct	link
AdaMergeX (Zhao et al., 2025b)	Composition Adaptation	Multiple	Multiple (Parameter modules optimized for different objectives)	Text Classification; QA; Summarization	MGSM; XCOPA; XNLI; XQuAD; XLSum	LLaMA-2-7B; XLM-R; T5-base, and mT5	link

ference cares more about how such reliability signals are translated into selective prediction. Therefore, when reliability estimation primarily serves as a decision policy for whether to answer, we put the methods under risk-controlled inference.

Moreover, for the research coverage, this survey primarily focuses on recent studies from major NLP and AI conferences and journals, including but not limited to ACL, EMNLP, NAACL, EACL, AACL, COLING, TACL, CL, NeurIPS, ICML, ICLR, AAAI, IJCAI, etc. Since LLM-based task adaptation has evolved rapidly in recent years, we give priority to works published between 2022 and 2026, while retaining earlier foundational studies when they provide essential background. In addition, we also include credible publications such as arXiv preprints and workshop papers, especially when they directly address low-resource task adaptation or offer transferable insights.

A.2 Low-Resource Tasks and Datasets

As a reference for future research on low-resource task adaptation, we summarize representative low-resource datasets and benchmarks widely applied in related work. It is noteworthy that some of these datasets are not inherently low-resource, but are commonly used in few-shot, limited-label, or scarce-supervision adaptation research. We organize them into three main categories as follows.

A.2.1 Text Classification and Semantic Understanding

Text classification and semantic understanding is one of the most common evaluation areas in existing low-resource studies. Representative tasks often focus on sentiment and emotion analysis, using datasets such as SST-2 (Socher et al., 2013), Rotten Tomatoes (Pang and Lee, 2005), and Emotion (Saravia et al., 2018), while also covering subjectivity

Table 3: Overview of representative reliable inference studies.

Work	Type	Method Input		Experimental Setting			Link
		Inference Strategy	External Resource	Task	Dataset	Foundation Model	
tRAG (Lee et al., 2025)	External-Resource-Augmented Inference	Reasoning with Retrieval	✓ (Evidence)	Information Retrieval	BEIR Benchmark	GPT-3.5-Turbo; Llama-3-8B; Llama-3.1-8B	N/A
Tool-Use Without Demonstrations (Luo et al., 2025c)	External-Resource-Augmented Inference	Reasoning with retrieval	✓ (Tools)	QA; Reasoning	TriviaQA; NQ-Open; PopQA; GSM8K	Llama-3-8B-Instruct; Llama-3-70B-Instruct	link
SCORE (Zhang et al., 2024c)	Reasoning-Enhanced Inference	Reasoning with Self-Correction	×/✓ (Optional Models)	QA; Reasoning	CommonsenseQA; QASC; RiddleSense; GSM8K; MATH	LLaMA-2-13B-chat; Gemma-7B-it	link
RetrievalPRM (Zhu et al., 2025a)	Reasoning-Enhanced Inference	Reasoning with Process Verification	✓ (Evidence)	Reasoning	GSM8K; MATH; OlympiadBench; Omni-MATH	Qwen-2.5-Math-7b-instruct	link
Tree-PLV (He et al., 2024)	Reasoning-Enhanced Inference	Reasoning with process verification	×	QA; Reasoning	CSQA, StrategyQA, GSM8K, and MATH	LLaMA2-7B; LLaMA2-13B; Mistral-7B; Mistral-7B-MetaMATH; WizardMath-7B	link
T^2 (Zhao et al., 2025c)	Reasoning-Enhanced Inference	Test-Time Scaling	×	QA	SQuAD; HotpotQA; BioASQ; NewsQA; GAOKAO; HQA; TriviaQA	Qwen2.5-32B-Instruct; GPT-4o; QwQ-32B-Preview	N/A
Shortest Majority Vote (Zeng et al., 2025b)	Reasoning-Enhanced Inference	Test-Time Scaling	×	QA; Reasoning	MATH-500; AIME; Omni-MATH; GPQA	DeepSeek-R1; DeepSeek-R1-Distill-Qwen-32b; DeepSeek-R1-Distill-Qwen-14b; DeepSeek-R1-Distill-Qwen-1.5b; QwQ; LIMO	N/A
Abstain with Multilingual Feedback (Feng et al., 2024a)	Risk-Controlled Inference	Inference with the Abstention Option	×	QA	M-MMLU; M-HellaSwag; Belebele	GPT-4; Aya-13B; ChatGPT	link
SConU (Wang et al., 2025g)	Risk-Controlled Inference	Inference with the abstention option	×	QA	MMLU; MMLU-Pro; MedMCQA; TriviaQA; CoQA	LLaMA-3.2-3B-Instruct; Qwen-2.5-3B-Instruct; Qwen-2-7B-Instruct; Qwen-2.5-7B-Instruct; OpenChat-3.5; LLaMA-3-8B-Instruct; LLaMA-3.1-8B-Instruct; LLaMA-2-13B-Chat; Vicuna-13B-v1.5; Qwen-2.5-14B-Instruct; Qwen-2.5-32B-Instruct	link

Table 4: Representative datasets and benchmarks for text classification and semantic understanding.

Task	Dataset / Benchmark	Scale / Coverage	Description
Sentiment Analysis	SST-2	9,613 sentence-level examples; 2 labels	Sentence-level sentiment benchmark with positive and negative labels.
Sentiment Analysis	SST-5	11,855 sentence-level examples; 5 labels	Fine-grained sentiment benchmark with labels from very negative to very positive.
Sentiment Analysis	Rotten Tomatoes	5,331 positive and 5,331 negative processed sentences	Movie-review sentiment benchmark with binary positive/negative labels.
Sentiment Analysis	Movie Review Data	2K polarity reviews; 10K subjective/objective sentences; rating-scale and sentence-level datasets	Cornell collection of movie-review datasets for sentiment and subjectivity experiments.
Question Classification	TREC	5.5K train / 500 test questions; 6 coarse and 50 fine labels	Benchmark for classifying questions according to expected answer type.
Subjectivity Classification	Subj	10K movie-review sentences; subjective/objective labels	Sentence classification benchmark for distinguishing subjective and objective text.
Emotion Classification	Emotion	20K English Twitter messages; 16K train / 2K validation / 2K test; 6 emotion labels	Text classification benchmark for identifying emotional categories in short text.
Hate Speech Detection	Hate Speech 18	10,568 Stormfront sentences; hate/non-hate labels	Text classification benchmark for identifying hate speech in Stormfront forum sentences.
Few-shot Text Classification Benchmark	RAFT	11 classification tasks; 50 examples per task; hidden test set	Real-world few-shot benchmark with hidden test labels.
Humanitarian Text Classification	HumSet	About 17K documents; 3 languages; 46 humanitarian projects	Domain-specific benchmark for humanitarian information extraction and multi-label classification in crisis-response analysis.
Textual Entailment	SICK	9,840 English sentence pairs; relatedness scores and entailment labels	Sentence-pair benchmark for semantic relatedness and entailment.
Relation Classification	SemEval-2010 Task 8	K=8/16 few-shot train/validation sets; 2,717 test samples; 19 relations	Relation classification benchmark for identifying semantic relations between marked nominals.
Semantic Similarity	STS-B	8,628 English sentence pairs; 5,749 train / 1,500 dev / 1,379 test; similarity scores from 0 to 5	GLUE task for measuring graded semantic similarity between two sentences.
Semantic Similarity	MRPC	About 5.8K English sentence pairs; news sources	Sentence-pair benchmark for paraphrase identification.

Table 5: Representative datasets and benchmarks for text generation, summarization, and translation.

Task	Dataset / Benchmark	Scale / Coverage	Description
Machine Translation	FLORES-200	3,001 sentences from 842 web articles; development / development-test / hidden test splits; 200+ languages	Multilingual MT evaluation benchmark for low-resource and many-to-many translation.
Code Generation	MultiPL-E	HumanEval and MBPP translated to 18 additional programming languages; unit-test-based evaluation	Multilingual code-generation benchmark for evaluating functional correctness across programming languages.
Summarization	DialogSum	13,460 dialogues; 12,460 train / 500 validation / 500 test; summaries and topics	Dialogue summarization dataset with real-life scenario dialogues, summaries, and topic annotations.
Summarization	SciTLDR	5.4K TLDRs over 3.2K computer science papers	Scientific extreme-summarization dataset with author-written and expert-derived TLDRs.

Table 6: Representative datasets and benchmarks for question answering and reasoning.

Task	Dataset / Benchmark	Scale / Coverage	Description
Question Answering	Few-Shot MRQA Benchmark	8 MRQA-style datasets; 0/16/32/64/128 training samples	Few-shot machine-reading QA setting where systems extract answer spans from contexts.
Question Answering	PopQA	14K entity-centric QA pairs; English; Wikidata triples with page views and relation types	Entity-centric long-tail factual QA benchmark grounded in Wikidata.
Question Answering	CommonsenseQA	12,247 multiple-choice questions	QA benchmark requiring commonsense knowledge and conceptual relations.
Question Answering	StrategyQA	2,780 yes/no questions; decompositions and evidence paragraphs	QA benchmark where questions require implicit reasoning strategies and supporting evidence.
Question Answering	MMLU-Pro	12K+ multiple-choice questions; 14 subjects; mostly 10 options per question	More challenging MMLU-style benchmark with expanded answer options and stronger reasoning difficulty.
Mathematical Reasoning	GSM8K (Grade School Math)	8.5K math word problems; 7.5K train / 1K test	Math word-problem benchmark with natural-language questions requiring multi-step arithmetic.
Mathematical Reasoning	MATH Dataset	12,500 competition math problems; step-by-step solutions	Challenging benchmark for multi-step mathematical problem solving.

classification (Pang and Lee, 2004), semantic similarity (Cer et al., 2017; Dolan and Brockett, 2005), textual entailment (Marelli et al., 2014), etc. In addition, several benchmarks, such as RAFT (Alex et al., 2021), provide a unified evaluation framework that covers diverse low-resource classification and language understanding tasks. A detailed comparison of related datasets and benchmarks is shown in Table 4.

A.2.2 Text Generation, Summarization, and Translation

Text generation, summarization, and translation also cover several important low-resource adaptation settings. Related datasets and benchmarks include FLORES-200 as a multilingual benchmark for machine translation (Costa-jussà et al., 2022), DialogSum and SciTLDR as summarization datasets from different domains and genres (Chen et al., 2021; Cachola et al., 2020), and MultiPL-E as a multilingual benchmark for evaluating code generation across programming languages (Cassano et al., 2022). Table 5 summarizes representative datasets and benchmarks in this category.

A.2.3 Question Answering and Reasoning

As popular tasks in LLM research, question answering and reasoning are also widely experimented in low-resource adaptation (see Table 6). Few-Shot MRQA evaluates few-shot extractive machine-reading comprehension (Ram et al., 2021), while PopQA focuses on open-domain factual QA over long-tail entities (Mallen et al., 2022). Reasoning-oriented QA benchmarks such as CommonsenseQA, StrategyQA and MMLU-Pro further test commonsense knowledge, implicit reasoning strategies, and multiple-choice reasoning (Talmor et al., 2019; Geva et al., 2021; Wang et al., 2024a). In addition, mathematical reasoning datasets, like Grade School Math and the MATH Dataset, are commonly applied to evaluate multi-step reasoning under limited supervision (Cobbe et al., 2021; Hendrycks et al., 2021).

A.3 New Mission for NLP Research: Low-Resource Task Adaptation Beyond Benchmarks

Low-resource task adaptation matters not only as a technical challenge, but also as a question of what kinds of real-world needs NLP research chooses to serve. In current benchmark-driven development, community attention is often disproportionately drawn to mainstream tasks and high-resource settings due to easier data collection, standard evaluation, and ready reward from academic and commercial communities. As a result, performance on many widely adopted benchmarks has already be-

come very strong, while numerous low-resource tasks with practical social value still remain under-explored.

In this sense, the importance of low-resource task adaptation goes beyond data efficiency alone. **It also concerns whether progress in NLP can reach needs that remain outside mainstream benchmark-centered development**. We therefore call for greater attention to the practical deployment of non-mainstream low-resource tasks, especially those languages and application scenarios that are systematically underserved. Such a shift is valuable not only for broadening the scope of LLMs, but also for aligning future research with socially urgent needs that may otherwise be overlooked due to limited market incentives and scarce annotation resources. This perspective is highly consistent with the United Nations' *Global Digital Compact* [5], which emphasizes closing digital divides, supporting underrepresented and vulnerable populations, promoting linguistic and cultural diversity, strengthening information integrity, and building AI capacity in developing and underserved contexts.

Appendix A.2 has introduced the common tasks in current low-resource studies. In this section, we highlight the **low-resource tasks that deserve greater attention** according to the suggestions from the *Global Digital Compact*.

A.3.1 Accessibility and Assistive Tasks for Vulnerable Groups

Accessibility and assistive tasks refer to LLM-based systems designed for persons with disabilities (e.g., blindness, deafness, autism, intellectual disability and obsessive-compulsive disorder), children, older adults, and other underserved users. These tasks include simplification, assistive dialogue, accessible information delivery, and age-appropriate or needs-aware interaction. Such tasks are often socially important but weakly represented in mainstream benchmarks. Table 7 presents related NLP tasks and corresponding datasets.

Existing resources have made encouraging progress in text accessibility, covering lexical simplification, sentence simplification, and complex word identification, such as TSAR-2022 (Saggion et al., 2022), PorSimplesSent (Leal et al., 2018), and the CWI 2018 shared task (Yimam et al., 2018). These resources are closely connected to practical accessibility needs, such as reducing reading barriers for dyslexic readers, low-literacy users, non-native speakers, and other users who may struggle with complex written information.

It is noteworthy that accessibility-oriented NLP extends beyond text simplification. Deaf and hard-of-hearing users, for example, are able to benefit from stronger support for sign-language-related research, while blind users may require accessible information delivery, visual assistance, and needs-aware dialogue systems. For example, the ChaLearn LAP 2021 signer-independent isolated sign language recognition challenge, built on the AUTSL dataset, provides a useful step toward multimodal accessibility evaluation (Sincan and Keles, 2020; Sincan et al., 2021). Nevertheless, compared with text-centered simplification benchmarks, standardized resources for broader assistive communication scenarios remain relatively limited. This gap suggests that low-resource task adaptation can further support NLP evaluation by extending benchmark coverage toward socially important applications that have not yet been sufficiently represented in mainstream benchmark design.

A.3.2 Low-Resource Languages, Dialects, and Localized Public-Service NLP

The Compact explicitly highlights linguistic and cultural diversity, local content, and inclusive access. This points to the need for stronger support for low-resource languages and dialects (e.g., Irish Gaelic, Breton, Hakka, Cantonese, and Hokkien), as well as multilingual and localized LLM-based systems for public services, community access, and region-specific communication needs. For low-resource task adaptation, these settings are challenging because linguistic resources are often unevenly distributed, while target applications may require sensitivity to local terminology, cultural norms, and public-service contexts.

Recent work shows that this demand is gradually taking shape in more concrete evaluation settings. Table 8 summarizes related datasets and benchmarks. For example, TICO-19 connects low-resource language methods with localized public-health communication by providing a COVID-19 translation benchmark across many lower-resourced languages (Anastasopoulos et al., 2020), while AmericasNLP 2025 focuses on machine translation for extremely low-resource Indigenous languages of the Americas (De Gibert et al., 2025). NusaX further expands this space to

[5] https://www.un.org/global-digital-compact

Table 7: Representative tasks for accessibility and assistive tasks for vulnerable groups.

Task	Dataset / Benchmark	Scale / Coverage	Description
Lexical Simplification	TSAR-2022 Shared Task on Multilingual Lexical Simplification	English, Spanish, and Brazilian Portuguese; about 1.1K instances overall; trial and test sets only.	Replaces complex words with simpler alternatives in context, supporting accessible reading for dyslexic, low-literacy, and language-learning users.
Text Simplification	PorSimplesSent (PorSimples Project)	Brazilian Portuguese; 4,888 original-to-simplified sentence pairs.	Portuguese sentence-pair corpus for readability assessment and simplification, derived from the PorSimples project.
Sign Language Recognition	ChaLearn LAP 2021: Signer Independent Isolated SLR Challenge (AUTSL)	RGB+D sign language videos; 36,302 video samples across 226 sign classes and 43 signers; user-independent train/development/test splits.	Benchmarks signer-independent isolated sign language recognition, supporting assistive communication for deaf and hard-of-hearing users.
Complex Word Identification	CWI 2018 Shared Task (Complex Word Identification)	English, German, Spanish, and French; about 62K labeled instances overall; train/development/test splits for English, German, and Spanish, plus a French test set.	Identifies words likely to be difficult for specific user groups, serving as a prerequisite task for accessible text simplification.

Table 8: Representative tasks for low-resource languages, dialects, and localized public-service NLP.

Task	Dataset / Benchmark	Scale / Coverage	Description
Machine Translation	TICO-19 Translation Benchmark	COVID-19 health documents; 3,071 sentences translated into 36 languages, including 26 lower-resource languages; 30 documents overall.	Supports access to COVID-19 information through health-domain translation in lower-resource languages.
Machine Translation	AmericasNLP 2025 Shared Task 1: Machine Translation for Indigenous Languages	Spanish and 13 Indigenous languages of the Americas; limited parallel data.	Evaluates truly low-resource Indigenous-language machine translation in both into- and from-Spanish directions.
Sentiment Analysis / Machine Translation	NusaX: Multilingual Parallel Sentiment Dataset for Indonesian Local Languages	12 languages, including 10 Indonesian local languages; 500 train / 100 development / 400 test instances per language.	Provides native-speaker-translated and verified resources for Indonesian local-language sentiment analysis and machine translation.
Arabic Dialect Identification	NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task	21 Arab countries and 100 provinces; 30,957 labeled tweets and 10M+ unlabeled tweets.	Supports fine-grained dialect identification at country and province levels using naturally occurring Twitter data.
Named Entity Recognition	MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition	20 African languages; publicly released train/development/test named entity recognition splits.	Provides a human-annotated named entity recognition benchmark for underrepresented African languages and cross-lingual transfer.
Semantic Relatedness	SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages	14 African and Asian languages; 26,257 sentence pairs.	Benchmarks sentence-pair semantic relatedness in underrepresented African and Asian languages.

Table 9: Representative tasks for online harm detection and intervention.

Task	Dataset / Benchmark	Scale / Coverage	Description
Persuasion Technique Detection	SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images	Memes; 687 train / 63 development / 200 test instances; 3 subtasks.	Benchmarks fine-grained persuasion technique detection in multimodal content.
Multimodal Hate / Sentiment Detection	CHiPSAL 2026 (ICWSM 2025 Shared Task)	Nepali Devanagari memes; 1,068 meme instances; 2 subtasks.	Targets hate speech detection and sentiment analysis in low-resource Nepali memes.
Misogyny Identification	AMI 2018: Automatic Misogyny Identification (EVALITA 2018)	Italian and English tweets; 4,000 train / 1,000 test tweets per language; 2 subtasks.	Supports misogyny identification, misogynistic behaviour categorization, and target classification in social media text.
Multi-task Hate Speech Detection	BLP 2025 Shared Task 1: Bangla Multi-Task Hate Speech Identification	Bangla YouTube comments; 3 subtasks; publicly available size not reported.	Detects hate type, target group, and joint hate, severity, and target labels in Bangla social media text.
Harm Detection in Text and Memes	MAHED2025 (ArabicNLP 2025)	Arabic and its varieties; 3 tasks; access upon request.	Covers hate and hope speech, offense, emotion, and hate detection, and hateful meme detection.
Persuasion Technique Detection	SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes	4 languages; 20 textual and 22 multimodal persuasion techniques; registration required.	Detects persuasion techniques in meme text and multimodal meme content.
Toxic Span Extraction	SemEval-2021 Task 5: Toxic Spans Detection	User-generated posts; 690 trial / 7,939 train / 2,000 test posts.	Identifies the exact spans responsible for toxicity in user-generated text.
Patronizing and Condescending Language Detection	SemEval-2022 Task 4: Patronizing and Condescending Language (PCL) Detection	English news paragraphs from 20 English-speaking countries; 10,637 paragraphs; access upon request.	Detects implicit condescension toward vulnerable communities in news text.
Multimodal Misogyny Detection	SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI)	Memes; 10,000 train / 1,000 test memes; academic access upon request.	Identifies misogyny in memes using both textual and visual signals.
Explainable Sexism Detection	SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS)	Reddit and Gab comments; 20,000 comments; 3 hierarchical subtasks.	Requires fine-grained and explainable sexism classification beyond binary detection.

Indonesian local languages through parallel sentiment and machine translation resources created with native-speaker translation and verification (Winata et al., 2023). Beyond translation and sentiment analysis, MasakhaNER 2.0 provides a human-annotated named entity recognition benchmark for 20 African languages (Adelani et al., 2022), and NADI 2020 offers fine-grained Arabic dialect identification at both country and province levels (Abdul-Mageed et al., 2020). Broader semantic understanding is also represented by semantic textual relatedness benchmarks for African and Asian languages (Ousidhoum et al., 2024). Taken together, these resources point to encouraging progress beyond general multilingual coverage, with increasing attention to public-service applications, Indigenous and local languages, dialectal variation, and information extraction. Nevertheless, real-world localized public-service scenarios remain comparatively underdeveloped, leaving substantial room for future task design and resource building.

A.3.3 Online Harm Detection and Intervention

The Compact places particular emphasis on online violence, hate speech, discrimination against underrepresented groups (e.g., racial/ethnic, sexual/gender, and religious minorities, as well as immigrants), gender-based abuse, cyberbullying, and child safety. These concerns naturally motivate low-resource LLM research on fine-grained harm detection, risk identification, reporting assistance, and safer intervention mechanisms, especially for vulnerable and underprotected communities. From the perspective of low-resource task adaptation, online harm is challenging because harmful behavior is highly context-dependent, rapidly evolving, and often expressed through community-specific language, implicit references, coded expressions, or multimodal content. As a result, models require adaptation not only to different languages and platforms, but also to finer harm taxonomies and local social contexts where labeled supervision may be limited or difficult to obtain.

Recent shared tasks suggest that this direction is increasingly reflected in low-resource and community-specific settings (see Table 9). For example, the Bangla Multi-Task Hate Speech Identification shared task frames harm detection as a multi-task problem over Bangla social media comments, while CHiPSAL extends this line to Nepali-only memes with multimodal hate-speech and sentiment subtasks, and MAHED2025 further broadens the scope to Arabic and its variants through hate/hope speech classification, multitask emotion/offensive/hate detection, and multimodal hateful meme detection. Earlier shared tasks such as AMI 2018 already established misogyny identification, misogynistic behaviour categorization, and target classification as benchmarkable problems in English and Italian social media (Fersini et al., 2018). More recent SemEval resources show that benchmark construction in this area is moving beyond coarse toxicity classification toward finer and more socially grounded formulations, including multilingual persuasion techniques in memes (Dimitrov et al., 2024), persuasion technique detection in texts and images (Dimitrov et al., 2021), explainable online sexism (Kirk et al., 2023), patronizing and condescending language toward vulnerable communities (Perez-Almendros et al., 2022), multimedia misogyny identification (Fersini et al., 2022), and toxic span extraction (Pavlopoulos et al., 2021). At the same time, these resources remain fragmented across languages, modalities, and harm taxonomies, and several emerging datasets are access-controlled rather than fully open, requiring registration or organizer approval before download.

A.3.4 Crisis Response and Public-Information Integrity Tasks

The Global Digital Compact highlights the importance of strengthening information integrity and ensuring reliable, timely, and multilingual information, especially in high-stakes public contexts. This concern is closely related to low-resource task adaptation: each new crisis, language, or domain can define a distinct target task, while labeled supervision is often scarce under such situations. Crisis-oriented and public-interest NLP therefore requires models that can adapt under limited supervision to tasks such as misinformation detection, fact-checking support, crisis-oriented summarization, multilingual event tracking, and other forms of public communication.

Existing shared tasks have begun to operationalize this direction through multilingual information integrity benchmarks (see Table 10). Fact-checked claim retrieval supports the reuse of verified evidence across languages and social-media contexts (Peng et al., 2025), while multilingual narrative characterization helps analyze how public events are framed in online news (Piskorski et al., 2025).

Table 10: Representative tasks for crisis response and public-information integrity.

Task	Dataset / Benchmark	Scale / Coverage	Description
Narrative Analysis	SemEval-2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News	Bulgarian, English, Hindi, Portuguese, and Russian online news; 2,500+ articles and 7,400+ entity mentions.	Identifies entity framing, narrative labels, and explanatory evidence in news about public crisis events.
Evasion Detection	SemEval-2026 Task 6: CLARITY – Unmasking Political Question Evasions	U.S. presidential interview question–answer pairs; Hugging Face size bucket of 1K–10K instances.	Detects clarity levels and fine-grained evasion strategies in political responses.
News Similarity	SemEval-2022 Task 8: Multilingual News Article Similarity	Multilingual news articles; nearly 10K article pairs across 18 language combinations.	Measures whether multilingual news articles describe the same event or share similar content.
Claim Retrieval	SemEval-2025 Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval	MultiClaim-based benchmark; 28K+ social media posts in 27 languages and 200K+ fact-checked claims.	Retrieves relevant previously fact-checked claims for social media posts in monolingual and cross-lingual settings.
Machine-Generated Text Detection	SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Machine-Generated Text Detection	9 languages; 76,863 machine-generated and 80,994 human-written texts in the training set.	Detects machine-generated, human-written, and mixed texts across domains and generators.

Table 11: Representative tasks for labor, education, and reskilling.

Task	Dataset / Benchmark	Scale / Coverage	Description
Skill Matching	TalentCLEF 2026 Task B: Job-Skill Matching with Skill Type Classification	English ESCO-based benchmark; 1,139 job-title queries and 5,358 skill elements in the test set.	Retrieves job-relevant skills and classifies them as core or contextual for reskilling support.
Readability Assessment	CLEAR Corpus	English reading passages for grades 3–12 readers; around 5K passage excerpts with readability scores and metadata.	Supports educational text selection and personalized reading-level assessment.
Skill Prediction	TalentCLEF 2025 Task B: Job Title-Based Skill Prediction	English ESCO-based benchmark; at least 5K train / 200 development / 500 test job titles with ESCO-normalized skills.	Predicts relevant professional skills from job titles to support career profiling and reskilling pathways.
Subject Tagging	SemEval-2025 Task 5: LLMs4Subjects	English and German technical records from TIBKAT; 123,589 records with the GND subject taxonomy.	Assigns structured subject labels to scientific and technical records for educational and research access.

Machine-generated text detection addresses the growing risk that automatically generated content may affect public information environments (Wang et al., 2024b), and multilingual news article similarity supports cross-lingual event tracking by identifying whether news articles describe the same or similar events (Chen et al., 2022). Recent work on political question evasion further extends this space toward public-facing communication, where models are expected to recognize strategic ambiguity or avoidance in political interviews (Thomas et al., 2026). Overall, these resources show encouraging progress toward evaluating information reliability in multilingual and high-stakes settings. Meanwhile, more direct crisis-response tasks, such as emergency information access, localized alert generation, and rapidly adaptive crisis summarization, remain relatively less represented, leaving substantial room for future low-resource research.

A.3.5 Tasks for Labor, Education, and Reskilling

Following the Global Digital Compact's emphasis on digital inclusion and AI capacity-building, this direction also concerns the impact of digitalization and automation on workforces, as well as the need for support in education, upskilling, and reskilling. This suggests an important role for low-resource LLMs in practical tasks such as instructional assistance, skill-oriented dialogue, career-transition support, and personalized educational tools for communities with limited resources. This priority is further supported by Falk and Tsoukalas (2026), suggesting that retraining and reemployment support can help mitigate the labor-market disruptions caused by AI-driven automation, although such measures may only partially offset the broader structural risks of over-automation.

From the perspective of low-resource task adaptation, these scenarios are challenging because educational and labor-market needs are often shaped by local curricula, occupation-specific skill definitions, and regional labor demands. Therefore, models adapted mainly to general-purpose benchmarks may not reliably support practical decisions such as selecting appropriate teaching materials, organizing educational resources, identifying job-relevant skills, or guiding reskilling pathways. Existing resources have begun to make these needs more measurable (see Table 11). For example,

the CLEAR Corpus supports readability assessment for educational text selection (Crossley et al., 2023), while LLMs4Subjects studies subject tagging for scientific and technical records in a bilingual setting (D'souza et al., 2025). In the labor-market domain, TalentCLEF 2025 and TalentCLEF 2026 operationalize skill intelligence through job-title-based skill prediction and job-skill matching with skill-type classification (Gasco et al., 2025; Gasco Sanchez et al., 2026). On the whole, these benchmarks suggest that labor and education tasks are not merely downstream applications of general NLP, but important low-resource adaptation settings where models must specialize to evolving needs, heterogeneous user groups, and resource-constrained decision contexts.

A.3.6 Discussion

Although the directions discussed above have been supported by several emerging datasets, they still occupy only a small fraction of mainstream NLP research. One reason is that many socially grounded low-resource tasks are inherently cross-disciplinary: defining the task, identifying meaningful labels, collecting data responsibly, and designing appropriate evaluation protocols often require close collaboration with domain experts in areas such as social psychology, law, medicine, education, and public policy. By contrast, tasks such as coding, mathematical reasoning, and other STEM-oriented benchmarks are often easier to standardize, scale, and evaluate within existing computational paradigms. This difference may partially explain why current LLM evaluation has placed relatively more emphasis on productivity-oriented capabilities than on socially situated needs.

Looking forward, we believe low-resource task adaptation can help broaden the agenda from *AI for productivity* toward *AI for social good*. This also creates an opportunity for researchers in universities and public research institutions to make distinctive contributions. Through interdisciplinary collaboration between academic departments, professional associations, nonprofit organizations, and social service agencies, they are well positioned to develop task definitions, datasets, evaluation protocols, and adaptation methods for domains that may not be immediately prioritized by commercial incentives.

Additionally, these underrepresented tasks also raise a practical question: **how can LLMs be rapidly and reliably adapted to the low-resource tasks discussed above?**

First, for tasks where task-specific annotations are extremely scarce, data-centric methods (§3) are often the most direct starting point, as the primary bottleneck lies in obtaining high-quality, diverse, and task-aligned supervision. In particular, evidence from instruction tuning suggests that, when the base model already possesses relevant knowledge, a small amount of carefully curated data can effectively guide model behavior toward the target objective (Zhou et al., 2023). In contrast, indiscriminately increasing the amount of pseudo data without quality control may instead introduce noise and degrade model performance (Ding et al., 2024).

Second, when a task has accumulated a usable amount of annotated data, the focus can further shift toward effective post-training (§4). In such settings, the key challenge is no longer only how to obtain supervision, but how to absorb limited supervision reliably without overfitting, unstable optimization, or catastrophic forgetting (Szep et al., 2026). Methods such as distillation and model merging can therefore play important roles, although some approaches, such as multi-task learning and meta-learning, may also introduce additional adaptation costs due to multiple supervision sources or repeated optimization procedures. In this regard, stability-oriented PEFT methods provide a lightweight alternative, reducing the computational burden of full-parameter adaptation while improving the efficiency of model updating under limited supervision (Patwa et al., 2024).

Finally, reliable inference (§5) provides a broadly applicable complement for low-resource deployment. For knowledge-intensive, long-tail, public-service, or high-stakes tasks, approaches such as retrieval augmentation, verification, test-time scaling, and selective prediction can help reduce unsupported predictions and improve decision reliability. These methods are not unique to low-resource NLP, but they become particularly important for tasks that require up-to-date knowledge, high factual accuracy, or carefully controlled risk (Zhang et al., 2025d; Wang et al., 2024c; Vu et al., 2024).

A.4 Research Community and Workshops in Low-Resource NLP

Addressing the multifaceted challenges of low-resource task adaptation requires collaborative community efforts. As data scarcity and real-world long-tail applications become increasingly

prominent, numerous workshops and shared task communities have emerged to tackle these bottlenecks. These communities serve as crucial hubs for curating domain-specific corpora, standardizing evaluation protocols, and fostering interdisciplinary collaboration. For instance, workshops like MRL and SIGTYP focus on cross-lingual transfer and representation learning as essential technical bridges. Simultaneously, regional initiatives such as AfricanNLP and AmericasNLP, alongside noisy-text communities like W-NUT and CALCS, have been instrumental in building foundational data infrastructure from the ground up for marginalized languages and complex real-world scenarios.

To provide a comprehensive resource for researchers, we compile an extensive overview of key NLP workshops dedicated to low-resource scenarios (see Table 12). However, a review of this landscape also reveals notable limitations in the current community focus. While traditional workshops have laid a solid foundation for data curation and basic NLP pipelines, there remains a discernible gap in directly addressing the modern LLM paradigm. Many existing communities are predominantly skewed toward resource creation and conventional cross-lingual transfer, with disproportionately less focus on advanced post-training adaptation, reasoning enhancement, or hallucination mitigation under extreme data scarcity. Although emerging workshops like LoResLM and UncertaiNLP have begun to bridge this divide, transitioning from basic resource building to trustworthy LLM adaptation requires more unified efforts. We hope this directory not only guides practitioners toward relevant collaborative platforms but also inspires future workshops to fill these critical methodological voids.

A.5 Surveys of Related Methodological Threads

This survey is organized around the central problem of low-resource task adaptation, rather than around individual methodological areas themselves. Beyond the taxonomy of low-resource task adaptation in Figure 3, we further present representative surveys on these related methodological threads in Figure 5. This list is not intended to replace specialized surveys in each area. Rather, it serves to clarify how our problem-driven organization relates to existing method-driven surveys. Although many of the methods covered by these surveys were not originally designed for low-resource task adaptation, they offer transferable insights for improving data construction, training efficiency and stability, and inference-time reliability under limited task supervision. By providing this mapping, we hope to motivate future work to adapt and validate these techniques in realistic low-resource task adaptation scenarios, contributing to the broader vision behind No Task Is an Island.

Table 12: Research communities and workshops in low-resource NLP.

Workshop / Track	Year(s)	Host Conference	Research Focus & Community Scope
LoResLM (Language Models for Low-Resource Languages)	2025–present (Annual)	COLING	Focuses specifically on adapting Large Language Models (LLMs) for low-resource languages, building dedicated benchmarks, and mitigating high-resource bias in generative AI.
UncertaiNLP (Uncertainty-Aware NLP)	2024–present (Annual)	EACL / EMNLP	Explores model reliability under extreme data scarcity. Focuses on open-ended discovery, hallucination mitigation, and confidence estimation when models are forced to operate under extreme data scarcity and out-of-domain.
MRL (Multilingual Representation Learning)	2021–present (Annual)	EMNLP	Focuses on Multilingual Representation Learning and cross-lingual transfer as the core technical bridge for low-resource tasks.
SIGTYP (Special Interest Group on Typology)	2019–present (Annual)	EACL / ACL / NAACL	Integrates linguistic typology into multilingual NLP. Focuses on leveraging structural features to improve cross-lingual transfer and joint learning for under-resourced languages.
LoResMT (Technologies for Low-Resource MT)	2018–present (Annual)	NAACL	A dedicated workshop for Low-Resource Machine Translation. Emphasizes the creation of parallel corpora, cross-lingual transfer, and shared tasks for low-resource and zero-resource language pairs.
WMT Shared Tasks (Low-Resource MT Track)	2006–present (Annual)	EMNLP	The premier benchmarking community for translation. Organizes competitive annual shared tasks specifically for low-resource machine translation to rigorously evaluate cross-lingual transfer and data augmentation methodologies.
CALCS (Linguistic Code-Switching)	2014–present (Periodic)	NAACL / EMNLP	Addresses the real-world phenomenon of code-mixing. Focuses on the extreme challenges of parsing informal social media text where low-resource native languages are heavily mixed with high-resource ones (e.g., English).
W-NUT (Workshop on Noisy User-generated Text)	2015–present (Except 2023)	EMNLP / COLING / ACL	Focuses on processing informal, noisy text (e.g., social media). Directly addresses data scarcity and code-mixing challenges that closely mirror low-resource NLP scenarios.
VarDial (Variation in Dialects and Languages)	2014–present (Annual)	COLING / NAACL / EACL	Focuses on computational methods and corpora for closely related languages, language varieties, and dialects, which often suffer from low-resource constraints.
AfricanNLP (African NLP)	2020–present (Annual)	ICLR / EACL	A grassroots regional community dedicated to the unique challenges of African languages. Emphasizes field data collection, speech recognition (ASR), and deploying ultra-compact models in resource-constrained environments.
AmericasNLP (NLP for Indigenous Languages of the Americas)	2021–present (Annual)	ACL / NAACL	Dedicated to indigenous languages of the Americas; explores machine translation, speech processing, and NLP tools for extremely data-scarce languages.
ArabicNLP (Formerly WANLP)	2014–present (Periodic)	ACL / EMNLP / EACL / COLING	While Modern Standard Arabic is resource-rich, this workshop extensively focuses on regional Arabic dialects and morphological complexities under severe low-resource constraints.
DravidianLangTech (Dravidian Language Technologies)	2021–present (Annual)	NAACL / EACL / ACL / RANLP	Focuses on developing speech and natural language processing technologies specifically for the highly morphologically complex and under-resourced Dravidian language family.
WILDRE (Workshop on Indian Language Data: Resources and Evaluation)	2012–present (Biennial)	LREC-COLING	Focuses on the creation, evaluation, and sharing of resources for Indian languages, specifically addressing under-resourced and code-mixed scenarios.
Field Matters (Workshop on Field-based NLP)	2022–present (Annual)	ACL / EACL / COLING	Addresses challenges in primary field data collection, digitization, and building tools for extremely under-documented and unwritten languages via speech and text processing.
ComputEL (Workshop on Computational Methods for Endangered Languages)	2014–present (Biennial)	ACL / EACL / ICLDR / ICLDC	Bridges the gap between computational linguists and field linguists to develop tools for endangered and extremely low-resource languages.
SIGUL Workshop (Special Interest Group on Under-resourced Langs)	2022–2024 (Transitioned to LT4All in 2025)	LREC-COLING	The central hub for standardizing datasets and tools for endangered and minority languages globally. (The community hosted the standalone LT4All conference in 2025).
LTEDI (Language Technology for Equality, Diversity & Inclusion)	2021–present (Annual)	EACL / ACL / RANLP / LDK	Promotes digital equality by developing NLP tools, hosting shared tasks, and building resources for marginalized, under-represented, and regional low-resource languages.

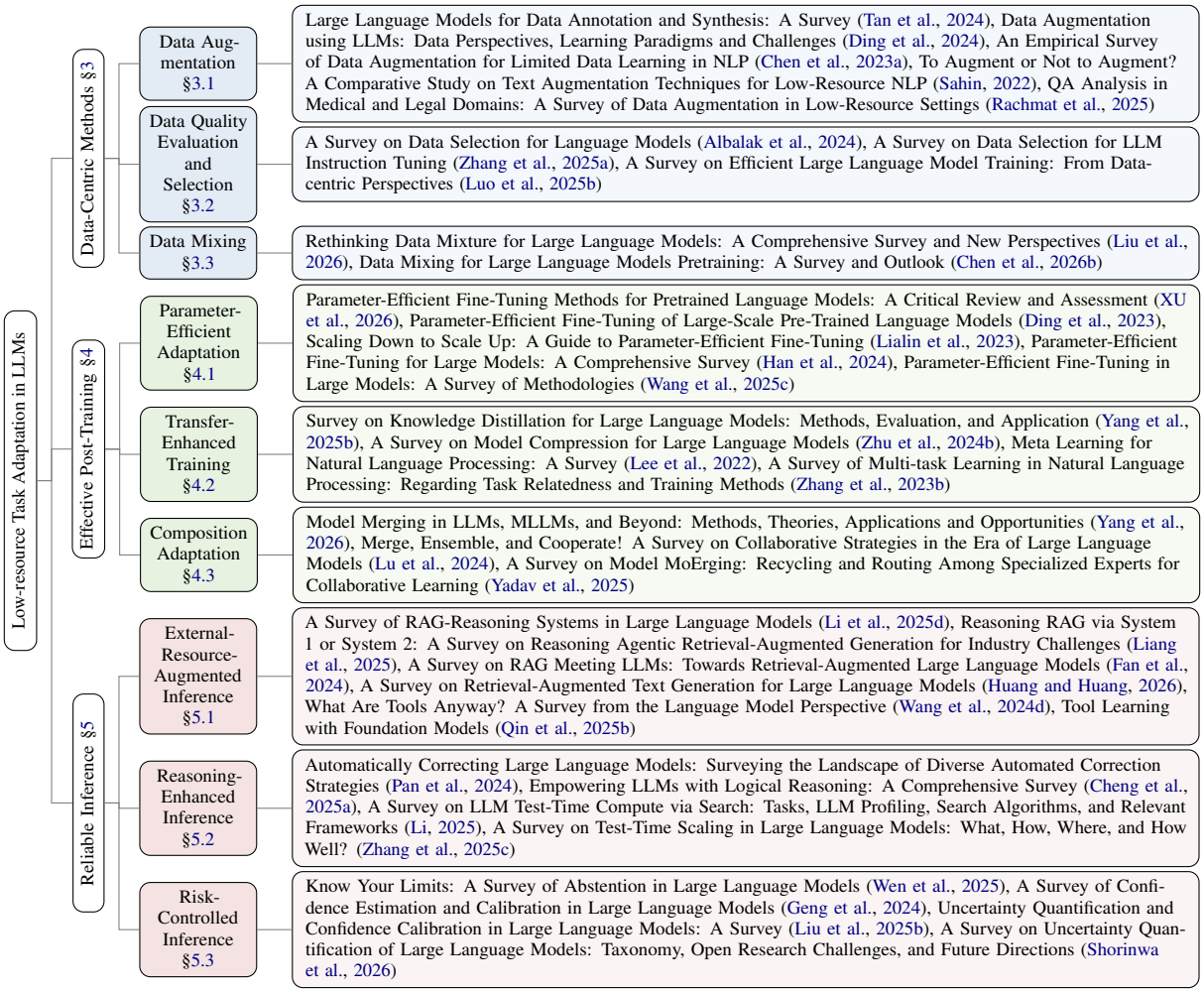

Figure 5: Representative surveys on methodological threads related to low-resource task adaptation. These surveys are grouped according to the three major components of our taxonomy: data-centric methods, effective post-training, and reliable inference. While many of these methodological threads were not originally developed specifically for low-resource task adaptation, they provide transferable insights for addressing scarce task-specific supervision.

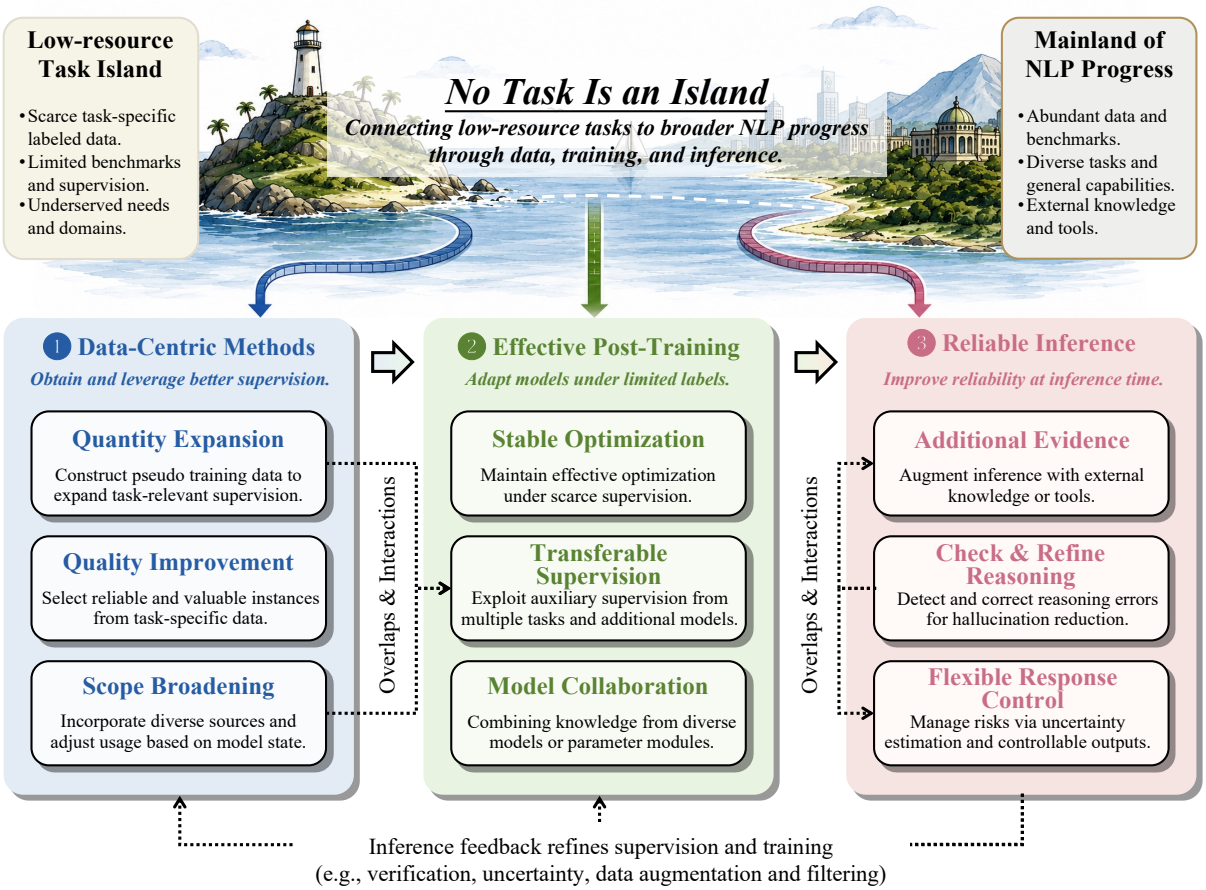

Figure 6: The scope of low-resource task adaptation and the overlaps between different research threads.

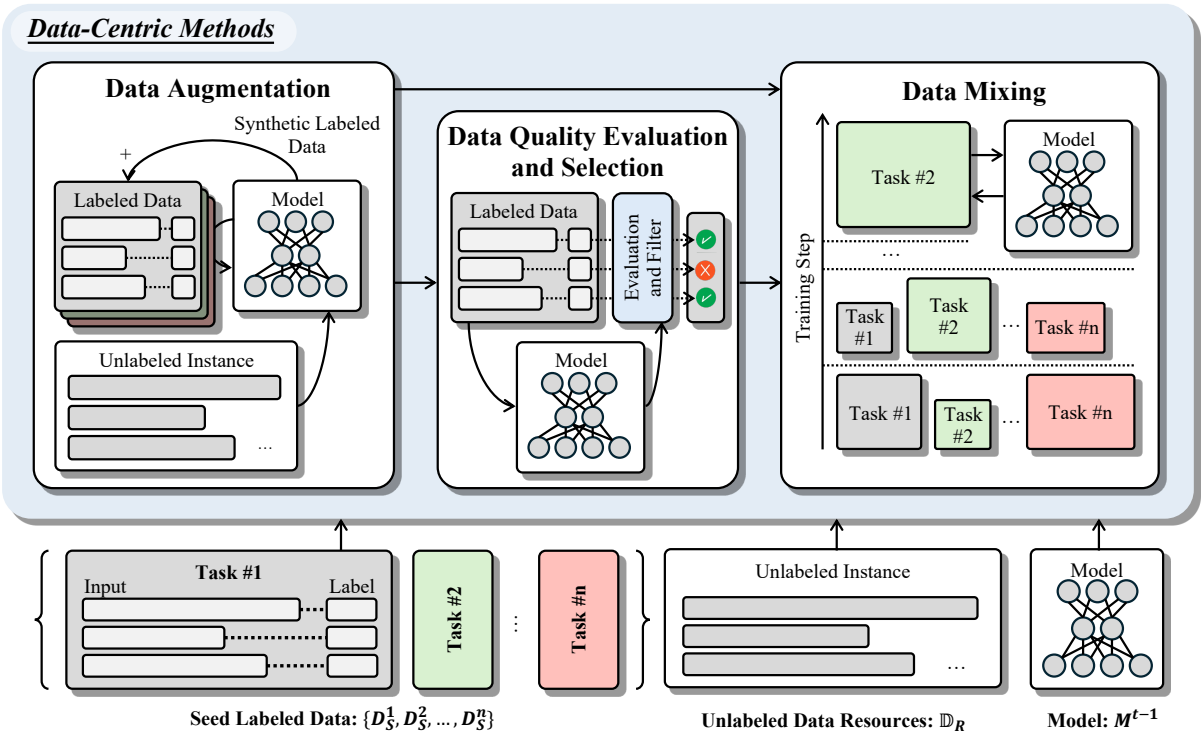

Figure 7: The overview of data-centric methods.

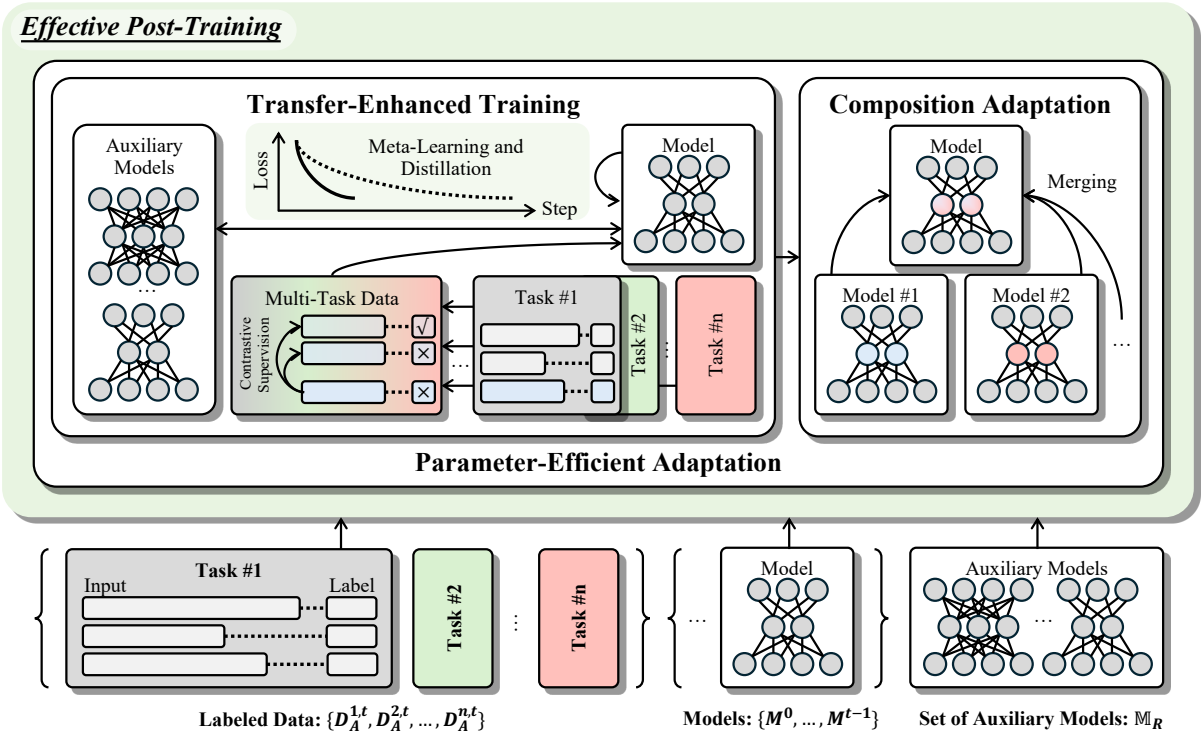

Figure 8: The overview of effective post-training.

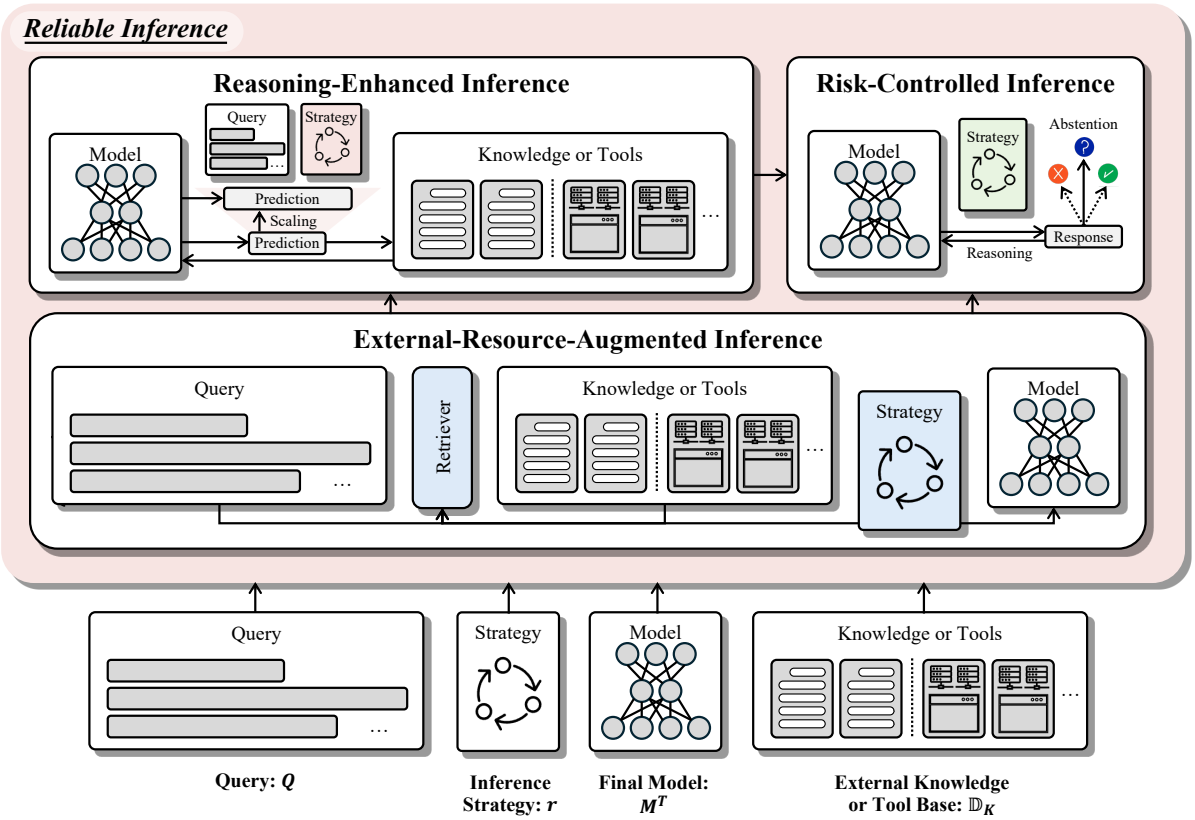

Figure 9: The overview of reliable inference.