Selective Reduction in Foundation Models: A Survey

Anonymous ACL submission

Abstract

Scaling has been the dominant paradigm behind recent advances in large language models and foundation models. However, a growing body of evidence shows that stronger performance does not always arise from adding more data, parameters, context, retrieved evidence, or reasoning computation. In many settings, models improve when redundant, noisy, misleading, or low-value information is selectively removed. This survey provides the first unified review of selective reduction methods from a performance-oriented perspective. We organize the literature into three stages of the model pipeline: input-side selection, model-internal selection, and inference-time selection. Across these stages, we show that seemingly diverse methods share common underlying mechanisms: they improve performance by densifying useful signals and reducing interference from irrelevant components. We further discuss the conditions under which selective reduction succeeds or fails, highlight limitations of current selection strategies, and outline open challenges toward turning selective reduction from a collection of local heuristics into a general design principle for future foundation models.

1 Introduction

Recent advances in foundation models have been strongly driven by scale (Kaplan et al., 2020; Hoffmann et al., 2022). Increasing the size of pretraining corpora, extending context length, scaling model parameters, incorporating more retrieved evidence, and allocating additional inference-time computation have all contributed to substantial performance gains (Wei et al., 2023; Lu et al., 2025). However, the relationship between scale and performance is not universally monotonic. In many practical scenarios, models are exposed to large amounts of information that are redundant, weakly relevant, noisy, or even misleading. Under such conditions, adding more information may not only yield diminishing returns, but can actively degrade model performance (Liu et al., 2024a; Shi et al., 2023; Albalak et al., 2024).

This phenomenon has motivated a rapidly growing body of work centered on selective reduction. In data-centric settings, several methods improve alignment and generalization by training on carefully selected subsets of high-quality supervision rather than the entire dataset (Zhou et al., 2023; Liu et al., 2024b, 2025a). In long-context modeling and retrieval-augmented generation (RAG), performance can be improved by filtering or compressing input evidence before it reaches the generator (Jiang et al., 2024; Wang et al., 2023b; Yan et al., 2024). Other approaches enhance downstream behavior by updating only selected model parameters or architectural components instead of the full model (Sharma et al., 2023; Zhang et al., 2025a; Naim et al., 2026). At inference time, improved predictions can also emerge from selecting only high-quality retrieved evidence, reasoning trajectories, or candidate outputs (Asai et al., 2024; Wang et al., 2023a; Toh et al., 2024).

Despite their shared intuition, these methods are typically studied within separate research paradigms, including data curation, prompt compression, sparse adaptation, retrieval filtering, and test-time reasoning. As a result, the broader conceptual connection among them remains insufficiently explored. The central claim of this survey is that many of these results are best understood not as isolated engineering tricks, but as evidence for a unified performance-oriented hypothesis: *less can be better when selection removes information or structure that would otherwise interfere with the model's use of relevant signal*.

Motivated by this perspective, we organize the literature by where the reduction occurs in the learning and inference pipeline. Input-side selection determines which supervision signals and contextual evidence should enter the model. Model-internal

selection determines which parameters, modules, or structural components should participate in adaptation and computation. Inference-time selection determines which retrieved evidence, reasoning paths, or candidate outputs should contribute to the final prediction. Across these settings, we focus on methods in which the reduction step itself improves accuracy, alignment, robustness, efficiency, or reasoning quality, and we use this unified framing to connect research directions that are otherwise studied independently.

Our survey is closely related to prior work on data selection (Albalak et al., 2024), which provides a comprehensive taxonomy of data selection methods across pretraining, instruction tuning, alignment, in-context learning, and task-specific fine-tuning. In contrast, we study data selection as a component of a broader selective-reduction framework for foundation models. Beyond training data selection, our survey also covers model-internal and inference-time selection and organizes these directions under a unified performance-oriented perspective. In summary, the main contributions of this survey are as follows:

- To the best of our knowledge, this is the first survey that unifies selective reduction methods under a common perspective, showing how selectively removing information or structure can improve model performance.
- We propose a unified taxonomy that organizes selective reduction methods into input-side, model-internal, and inference-time selection, enabling systematic comparison across previously disconnected research directions.
- We review the major empirical findings, underlying mechanisms, and open challenges in this emerging area, and discuss future directions for improving foundation models.

2 Why Selective Reduction Matters?

Although the surveyed methods differ substantially in mechanism, stage of application, and granularity, many of them can be understood through several shared principles.

Increasing Signal Density. A central idea underlying selective reduction is that model performance depends more on the density of useful signal than on the sheer quantity of available information. In instruction tuning, low-quality or weakly informative examples can dilute optimization signals and reduce training efficiency. This observation helps explain why small but carefully curated datasets such as LIMA, as well as automatically selected datasets such as DEITA and SelectIT, can match or even outperform substantially larger corpora (Zhou et al., 2023; Liu et al., 2024b, 2025a). A similar phenomenon appears in long-context modeling and prompt compression. When only a small portion of the context is truly relevant to the task, removing redundant or distracting content can make the remaining evidence easier for the model to utilize effectively (Jiang et al., 2024; Weston and Sukhbaatar, 2023).

Reducing Information Interference. Another key principle is that irrelevant information is often actively harmful rather than merely unhelpful. Retrieved documents may conflict with the model's parametric knowledge, weak reasoning trajectories may dominate majority voting procedures, and indiscriminate parameter updates may overwrite useful general knowledge during task adaptation. Many selective reduction methods improve performance by suppressing components that interfere with more reliable evidence or computation paths. This pattern can be observed in retrieval-correction methods such as CRAG and Self-RAG, selective parameter-updating strategies, and reasoning-path verification frameworks (Yan et al., 2024; Asai et al., 2024; Zhang et al., 2024; Toh et al., 2024).

Discussion. Taken together, these observations suggest that selective reduction is not merely a strategy to reduce computational cost or model size. More fundamentally, it serves as a mechanism to restructure information flow within foundation models, making task-relevant signals more prominent while limiting the influence of noisy, redundant, or conflicting information. This perspective provides a unified lens for understanding a broad range of methods, including training-sample selection, context filtering, structural sparsification, retrieval selection, and answer verification.

At the same time, the effectiveness of selective reduction is inherently conditional. Reduction is less likely to help when useful evidence is already sparse, when rare or long-tail cases are critical, or when the selection mechanism cannot reliably distinguish signal from noise. Consequently, the principle that "less can be better" should not be interpreted as a universal preference for smaller

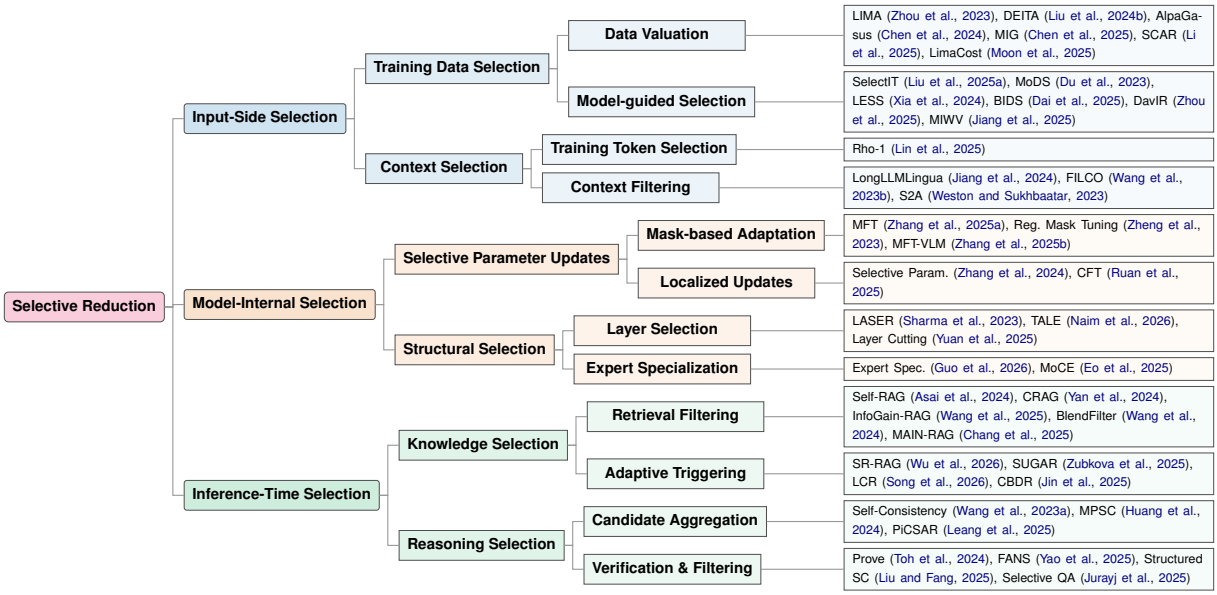

Figure 1: Taxonomy of performance-oriented selection. We organize methods by *where* the subtraction occurs: input-side selection removes low-value training samples or input evidence; model-internal selection activates or updates only task-relevant model parameters or model structure; inference-time selection filters retrieved evidence or reasoning candidates. In each leaf, we list representative methods reviewed in this survey.

datasets, shorter contexts, or more compact models. Rather, it reflects a conditional claim that performance can improve when reduction selectively removes redundancy and interference while preserving task-relevant information.

3 Input-Side Selective Reduction

Input-side selective reduction concerns what information enters the model before or during learning. Rather than assuming that more data always improves performance, these methods attempt to identify which supervision signals, training distributions, or contextual evidence are actually beneficial for the target objective. Existing approaches differ substantially in granularity, yet they share a common goal: improving model behavior by selectively preserving high-value information while suppressing redundancy, noise, and interference.

Instruction and Supervision Selection. The clearest evidence for the "less can be better" hypothesis appears in instruction tuning and alignment. LIMA demonstrates that a 65B LLaMA model fine-tuned on only 1,000 carefully curated examples can exhibit surprisingly strong alignment behavior, suggesting that much of the required knowledge is already acquired during pretraining and that post-training quality depends heavily on the choice of demonstrations rather than dataset scale alone (Zhou et al., 2023). DEITA further sharpens this observation by ranking instruction-response pairs according to complexity, quality, and diversity, showing that much smaller subsets can preserve or even improve alignment quality relative to substantially larger corpora (Liu et al., 2024b). Similarly, AlpaGasus shows that filtering noisy instruction data into a compact high-quality subset can outperform the original full dataset (Chen et al., 2024).

Subsequent work increasingly shifted from generic quality filtering toward model-aware data valuation. SelectIT estimates which examples are most beneficial for a target model through uncertainty-aware self-reflection (Liu et al., 2025a). MoDS organizes selection around quality, coverage, and necessity, explicitly asking which instructions remain useful after an initial tuning stage (Du et al., 2023). LESS formulates data selection as task-targeted subtraction by estimating example influence through low-rank gradient similarity, arguing that examples are valuable only when they induce updates aligned with desired downstream capabilities (Xia et al., 2024). Related approaches, such as BIDS and DavIR, continue this model-dependent perspective by balancing influence across capability groups or by estimating how learnable and reward-aligned each example is for the downstream objective (Dai et al., 2025; Zhou et al., 2025).

Another line of work studies which dataset

properties should remain preserved after reduction. MIG emphasizes information gain and global semantic coverage rather than relying solely on local heuristics (Chen et al., 2025). LimaCost estimates the value of training examples relative to the gradient contribution of a strong reference dataset (Moon et al., 2025). SCAR shows that stylistic consistency across high-quality examples can improve the effectiveness of supervision, suggesting that coherent response styles may make learning signals easier for the model to internalize (Li et al., 2025). Importance-aware data selection reaches a similar conclusion from another perspective, arguing that the most valuable examples for a given model are often those that expose its current weaknesses rather than those that appear globally clean or diverse (Jiang et al., 2025).

At the same time, recent analyses caution against interpreting selective reduction as universally beneficial. Nayak et al. show that the effectiveness of targeted instruction selection depends strongly on the underlying data representation, the selection algorithm, and especially the data budget (Nayak et al., 2026). In many cases, the largest gains occur in low-budget regimes and diminish as more data is retained. These findings refine the broader thesis of selective reduction: performance gains arise not from reduction itself, but from removing the right forms of redundancy and interference under identifiable conditions.

Pretraining Corpus and Data Mixture Selection. Selective reduction also operates at the scale of large pretraining corpora. Instead of selecting individual instruction examples, these methods focus on how entire domains, document collections, or semantic distributions should be filtered and combined under fixed training budgets. Data deduplication provides one of the earliest examples of performance-oriented corpus reduction. Lee et al. show that removing exact and near-duplicate documents reduces memorization and train-test overlap while maintaining or even improving language modeling performance (Lee et al., 2022). Beyond deduplication, Thrush et al. (2025) select pretraining domains using perplexity–benchmark correlations across existing LLMs, showing that model loss patterns can identify data domains that are more useful for downstream performance. Recent methods further optimize the composition of pretraining mixtures themselves. DoReMi (Xie et al., 2023) applies group distributionally robust optimization to reweight pretraining domains, demonstrating that selective domain balancing can improve downstream accuracy while reducing the number of training steps required to achieve a target level of performance. RegMix (Liu et al., 2025b) further reformulates pretraining mixture design as a regression problem, using proxy models trained on different mixtures to predict the performance of unseen combinations, thereby automating the discovery of effective data distributions.

More recent work emphasizes semantic quality and long-horizon diversity preservation. FineWeb-Edu (Penedo et al., 2024) filters web-scale corpora using an educational-quality classifier, showing that selecting a smaller subset of educationally valuable documents substantially improves reasoning- and knowledge-oriented benchmarks. Nemotron-CC (Su et al., 2025) highlights the trade-off between aggressive quality filtering and coverage preservation. Rather than simply discarding low-quality documents, it combines classifier ensembling, synthetic rephrasing, and reduced reliance on heuristic filters to preserve a broader set of unique high-quality tokens. Nemotron-CLIMB (Diao et al., 2025) further extends this idea by replacing manually defined domains with automatically discovered semantic clusters, demonstrating that selective semantic mixture construction can improve pretraining under fixed token budgets.

Collectively, these methods suggest that pretraining quality depends not only on dataset scale but also on how effectively the retained corpus allocates modeling capacity across domains, semantic regions, and informational content. The resulting perspective reframes corpus construction as a selective optimization problem rather than a purely scaling-driven process.

Context and Token Selection. Selective reduction also occurs below the level of entire training examples or document collections. Rather than deciding which samples to retain, these methods determine which tokens, passages, or contextual evidence should remain visible to the model during training or inference. During training, Rho-1 demonstrates that not all pretraining tokens contribute equally to learning (Lin et al., 2025). By selectively applying training loss only to informative tokens, the method improves mathematical and general downstream performance while using only a fraction of the original token stream. This result is important because it shows that selective

reduction can operate at the level of fine-grained micro-signals within each example, rather than only through sample-level filtering.

At inference time, long-context modeling and retrieval-augmented generation provide especially strong evidence that more context can sometimes degrade performance. LongLLMLingua improves long-context reasoning by compressing prompts in a query-aware manner, preserving task-relevant tokens while removing distracting or redundant content (Jiang et al., 2024). FILCO and System 2 Attention similarly improve downstream accuracy by filtering, restructuring, or regenerating contextual evidence so that irrelevant passages exert less influence on the final prediction (Wang et al., 2023b; Weston and Sukhbaatar, 2023). These methods collectively suggest that context quality often matters more than context quantity, and that removing noisy or weakly relevant evidence can directly improve the quality of reasoning and answer generation.

4 Model-Internal Selective Reduction

Model-internal selective reduction investigates whether better downstream behavior can emerge by activating, updating, or preserving only part of a foundation model. Unlike input-side selection, these methods operate on the model itself by selectively restricting parameter updates, computation pathways, or structural adaptation.

Selective Parameter Updates. One major line of work starts from the observation that many downstream capabilities already exist inside pretrained weights, and that full fine-tuning may disturb these capabilities too broadly. Mask Fine-Tuning (MFT) makes this intuition explicit by learning sparse masks over existing parameters rather than uniformly updating all weights (Zhang et al., 2025a). Regularized Mask Tuning reaches a similar conclusion in vision-language transfer, arguing that effective adaptation can emerge from uncovering task-relevant subnetworks instead of indiscriminately rewriting the full model (Zheng et al., 2023). Related evidence from broader vision-language adaptation further suggests that mask-based methods can outperform standard low-rank adaptation by more effectively reusing structures already encoded within pretrained weights (Zhang et al., 2025b).

Selective parameter update methods extend this idea by explicitly restricting optimization to sparse task-relevant subsets of the network. Zhang et al. show that updating only carefully selected parameters can incorporate new task knowledge while reducing damage to broad general-purpose capabilities (Zhang et al., 2024). Critical Token Fine-tuning introduces a more localized form of selective adaptation for reasoning tasks. Instead of propagating gradients uniformly across all tokens, it identifies counterfactually important reasoning tokens and selectively concentrates supervision around them, achieving stronger mathematical reasoning performance while updating fewer supervised positions (Ruan et al., 2025).

Structural Selection. A second line of work performs selective reduction at the level of larger internal model structures, such as layers, ranks, or architectural components. These methods suggest that some internal computations are not merely redundant but can actively degrade task performance. LASER demonstrates that selectively reducing the rank of specific layers can improve reasoning performance even after training has already completed (Sharma et al., 2023). This result is particularly striking because it implies that certain higher-order representational components may interfere with downstream reasoning rather than support it. TALE makes a related argument at inference time by dynamically removing task-irrelevant layers, showing that the resulting task-adapted architecture can match or even outperform the original full model (Naim et al., 2026). Similarly, work on layer pruning and layer-cutoff reports that smaller effective architectures can outperform intact models on specific prompt-based tasks (Yuan et al., 2025).

Compared with selective parameter updates, structural selection operates at a coarser level of abstraction. Rather than deciding which weights should change, these approaches determine which larger computational components should remain active at all. This distinction is important because the source of interference may not originate solely from parameter updates, but also from unnecessary representational transformations or reasoning pathways introduced by deeper layers. In this sense, structural selection reframes model adaptation as a problem of computational routing rather than only parameter optimization.

Discussion. Taken together, model-internal selective reduction supports a strong version of the central thesis of this survey: better performance can sometimes emerge not from increasing model par-

Table 1: Taxonomy of performance-oriented selection. **Blue**, **orange**, and **green** mark the three pipeline stages. **T/I** indicate whether the reduction primarily occurs at **T**raining or **I**nference time.

Sub-category	What is reduced	Selection signal	T	I	Why less is better	Risk
Input-Side Selection						
Data valuation	training samples	quality, diversity	✓	✗	noisy or redundant data dilutes supervision signal	coverage gaps
Model-guided selection	training samples	influence, uncertainty	✓	✗	model-aware selection targets current weaknesses	selector bias
Context selection	input evidence	relevance, compression	✓	✓	distracting tokens actively interfere with prediction	evidence loss
Model-Internal Selection						
Mask-based adaptation	model parameters	mask salience	✓	✗	broad updates damage useful pretrained structure	mask drift
Localized updates	model parameters	update criticality	✓	✗	narrower updates preserve general knowledge	underfitting
Structural selection	model structure	task utility	✓	✓	inactive or harmful components degrade quality	structure mismatch
Inference-Time Selection						
Retrieval filtering	retrieved evidence	utility, confidence	✗	✓	irrelevant evidence conflicts with parametric knowledge	filter errors
Adaptive retrieval triggering	retrieval choices	uncertainty, confidence	✗	✓	unnecessary retrieval injects noise	retrieval misses
Candidate aggregation	reasoning candidates	agreement, confidence	✗	✓	wrong candidates dominate majority voting	vote bias
Answer verification	reasoning candidates	external checks	✗	✓	fluent but incorrect chains contaminate output	verifier errors

ticipation, but from restricting it. Pretrained models do not distribute useful capabilities uniformly across parameters, layers, or reasoning pathways, and allowing the entire network to participate indiscriminately may introduce harmful interference, catastrophic forgetting, or inefficient computation.

At the same time, the success of these methods depends critically on the quality of the selection mechanism itself. Parameter masking, sparse updating, token-level supervision selection, and structural pruning all require reliable signals for identifying which components are beneficial and which are harmful. When the selector fails to distinguish useful structure from essential capability, selective reduction can easily degrade performance rather than improve it. Consequently, the key challenge is not merely to use fewer parameters or layers, but to identify the appropriate unit and selection criterion for the dominant source of interference in a given task or adaptation setting.

5 Inference-Time Selective Reduction

Inference-time selective reduction concerns what information or reasoning trajectories should survive until the final prediction. What changes is the policy for selecting which retrieved evidence, intermediate reasoning paths, or generated candidates the model should rely on during inference.

Knowledge Selection. Retrieval-augmented generation (RAG) systems provide one of the clearest demonstrations of why more information can sometimes hurt performance. Retrieved evidence frequently contains irrelevant, conflicting, redundant, or low-quality content. As a result, the practical challenge is not simply to retrieve more documents, but to determine whether retrieval is necessary and which retrieved evidence should remain available to the generator. Self-RAG internalizes part of this decision process by teaching the model to retrieve, critique, and utilize evidence selectively rather than consuming a fixed number of passages for every query (Asai et al., 2024). CRAG introduces an explicit retrieval evaluator together with a corrective retrieval pathway, showing that generation quality improves when the system can detect poor retrieval and repair or filter it before answer generation (Yan et al., 2024). InfoGain-RAG sharpens this perspective further by explicitly estimating how much each retrieved document improves downstream answer generation, effectively reframing retrieval selection as a direct utility estimation problem for the generator (Wang et al., 2025). Related methods differ substantially in implementation but converge on the same underlying principle. BlendFilter, MAIN-RAG, knowledge checking, and DRAGIN all improve retrieval-augmented generation by filtering evidence according to usefulness rather than allowing all retrieved information to compete equally during inference (Wang et al., 2024; Chang et al., 2025; Zeng et al., 2025; Su et al., 2024).

Another branch of work shifts the selection decision earlier, before or during retrieval itself. Self-Routing RAG frames selective retrieval as a knowledge-source routing problem, enabling the model to dynamically decide whether to retrieve external evidence or rely on its internal parametric knowledge (Wu et al., 2026). SUGAR similarly uses contextual confidence to trigger retrieval only when additional evidence appears necessary (Zubkova et al., 2025). Confidence-based reranking and post-retrieval calibration methods extend the same idea after retrieval, showing that model confidence signals can help identify which evidence is sufficiently reliable to trust during generation (Song et al., 2026; Jin et al., 2025). Across these approaches, the recurring insight is that retrieval effectiveness depends critically on selective suppression. Smaller, carefully filtered retrieval sets often produce more accurate outputs than larger collections of noisy evidence.

Reasoning and Candidate Selection. Selective reduction can also operate after the model has already generated reasoning candidates. Self-Consistency provides the canonical example. Instead of trusting a single greedy chain of thought, the method samples multiple reasoning trajectories and selects the answer that is most consistently supported across them (Wang et al., 2023a). Multi-Perspective Self-Consistency extends this idea by comparing candidates across multiple generated views, including solutions, specifications, and tests, thereby improving quality through structured selection among alternatives rather than through larger base models alone (Huang et al., 2024). PiCSAR further shows that even within best-of-N reasoning, improved candidate scoring can simultaneously reduce the number of sampled trajectories and improve final answer quality (Leang et al., 2025).

More recent work emphasizes selective reasoning under explicit computational constraints. Sequential reward filtering improves over standard best-of-N inference by carrying forward only high-reward generations, concentrating computation on promising candidates while suppressing weaker reasoning trajectories early in the process (Yu et al., 2025). Budget-aware test-time scaling through discriminative verification reaches a similar conclusion: under fixed inference budgets, better filtering and verification strategies can outperform substantially more expensive generative search procedures (Montgomery et al., 2025). Verification-based approaches make the subtraction mechanism even more explicit. Programs-as-Verifiers filters candidate mathematical solutions by checking whether translated programs agree with the original natural-language reasoning process (Toh et al., 2024). FANS further strengthens this idea by using formal verification with Lean4, selecting only answers that satisfy stronger symbolic-consistency constraints (Yao et al., 2025). Structured self-consistency for mathematics similarly shows that gains arise not only from selecting the most common answer, but also from eliminating intermediate reasoning traces that are internally inconsistent (Liu and Fang, 2025).

A complementary line of work asks whether parts of the reasoning trace itself should be removed. Think Clearly reports that pruning redundant reasoning tokens can improve downstream reasoning accuracy, particularly when long chains of thought become diffuse or repetitive (Choi et al., 2025). Related studies show that suppressing low-value verification or repetitive segments can improve performance, whereas indiscriminately pruning essential reasoning steps degrades performance (Zhao et al., 2025). In a similar spirit, Self-Verification Dilemma argues that reflective checking is frequently overused and often merely confirmatory; suppressing unnecessary self-verification steps can therefore preserve or even improve accuracy while reducing inference cost (Long et al., 2026). Selective question answering introduces another form of inference-time reduction through abstention and confidence thresholding. Instead of forcing the model to answer every query, these methods improve reliability by selectively refusing uncertain predictions, suggesting that answering fewer questions more confidently can sometimes yield better overall decision quality (Jurayj et al., 2025).

Taken together, this literature highlights an important refinement of the broader selective reduction thesis. Inference-time improvement does not necessarily require more retrieval, more reasoning steps, or more search trajectories. What often matters more is which candidates are retained, which are verified, and which are removed before they can interfere with the final decision process.

6 Open Problems

Despite rapid progress, selective reduction in foundation models remains far from fully understood.

Table 2: Representative "less can be better" evidence. **Kept** shows the reduced subset or setting; **Gain** reports improvement over the corresponding full or unselected baseline from the original paper. **Blue**, **orange**, and **green** mark the three pipeline stages. **T/I** indicate whether the reduction primarily occurs at **T**raining or **I**nference time.

Method	What is reduced	Selection signal	T	I	Kept	Gain vs. full baseline
LIMA (Zhou et al., 2023)	training samples	expert curation	✓	✗	2% dataset	+27.0% net win rate (LLaMA-65B)
AlpaGasus (Chen et al., 2024)	training samples	quality filter	✓	✗	17% dataset	+18.8% net win rate (LLaMA-13B)
DEITA (Liu et al., 2024b)	training samples	complexity, quality, diversity	✓	✗	3% dataset	+4.0% avg. acc. (Mistral-7B)
LESS (Xia et al., 2024)	training samples	influence	✓	✗	5% dataset	+2.3% avg. acc. (Mistral-7B)
Rho-1 (Lin et al., 2025)	training tokens	token utility	✓	✗	40% tokens	+16.5% avg. math acc. (TinyLlama-1.1B)
LongLLMLingua (Jiang et al., 2024)	prompt tokens	contrastive perplexity	✗	✓	4× compressed prompt	+11.6% NQ 20th acc. (GPT-3.5-Turbo)
MFT (Zhang et al., 2025a)	model parameters	learnable mask score	✓	✗	90% parameters	+6.0% IF-Eval score vs. SFT (LLaMA3.1-8B)
S-MFT (Zhang et al., 2025b)	model parameters	learnable mask score	✓	✗	∼97% parameters	+1.5% MMMU acc. vs. SFT (Qwen2.5-0.5B)
R-MT (Zheng et al., 2023)	model parameters	regularized mask gradient	✓	✗	97% parameters	+18.7% avg. acc. vs. zero-shot (CLIP)
CFT (Ruan et al., 2025)	training tokens	counterfactual criticality	✓	✗	<12% tokens	+6.4% avg. math acc. vs. SFT (Qwen2.5-7B)
LASER (Sharma et al., 2023)	model parameters	singular value magnitude	✗	✓	∼1% matrix rank	+10.9% CounterFact acc. (GPT-J 6B)
TELL-TALE (Naim et al., 2026)	transformer layers	validation accuracy	✗	✓	adaptive layer subset	+22.0% GSM8K-Hard acc. (LLaMA3.1-8B)
Self-RAG (Asai et al., 2024)	retrieved evidence	reflection tokens	✓	✓	adaptive passages	+16.7% PopQA acc. (LLaMA2-7B)
SR-RAG (Wu et al., 2026)	retrieved evidence	LLM uncertainty & kNN policy	✓	✓	64% retrieval	+10.5% avg. acc. (LLaMA2-7B-Chat)
InfoGain-RAG (Wang et al., 2025)	retrieved evidence	information gain	✗	✓	top-4 subset vs. top-100	+17.8% NQ acc. (Qwen2.5-72B)
Self-Consistency (Wang et al., 2023a)	reasoning candidates	agreement	✗	✓	majority answer	+17.9% GSM8K acc. vs. greedy (GPT-3)
Prove (Toh et al., 2024)	reasoning candidates	external checks	✗	✓	verified subset	+18.2% GSM8K acc. vs. vanilla (LLaMA2-7B-Chat)
FANS (Yao et al., 2025)	reasoning candidates	external checks	✗	✓	verified subset	+8.33% AMC-23 acc. vs. reward (DeepSeek-Math-7B)

A central open problem is how to coordinate selection across stages. Most existing methods optimize a single local decision, such as selecting training samples, filtering retrieved documents, pruning reasoning trajectories, or restricting parameter updates. Yet these decisions are likely interdependent: a model trained on selectively curated supervision may require a different retrieval strategy at inference time.

Another challenge is to move from heuristic selection signals to principled, end-task-grounded mechanisms. Current approaches often rely on uncertainty, confidence, influence, or information gain as proxies for usefulness, but the field still lacks a unified account of when such signals are reliable. Stronger causal, optimization-based, or information-theoretic analyses could make selective reduction more predictable and less dependent on dataset-specific heuristics.

Finally, selective systems must remain robust under distribution shift and be evaluated with standardized protocols. Information that is usually redundant may become important under changing distributions, especially in retrieval-augmented generation, multimodal evidence selection, and dynamic expert routing. Future evaluation should therefore measure not only whether a reduced system outperforms a full baseline, but also coverage preservation, calibration, robustness, computational efficiency, and stability across task families.

7 Conclusion

This survey examined selective reduction as a performance-oriented principle for foundation models. Across input selection, model adaptation, and inference reasoning, prior work shows that removing redundant, noisy, or conflicting components can improve behavior by preserving useful signal and reducing interference. The principle is conditional rather than universal: reduction helps when interference is substantial, but can fail when useful evidence is sparse, or the selector is unreliable. As foundation models continue to scale, future progress may depend not only on increasing capacity but also on learning what information, structure, and computation should be excluded.

Limitations

This survey is intentionally focused on performance-improving selection. That focus means some adjacent lines of work are discussed only briefly or not at all, and the resulting taxonomy should be read as a working synthesis of this specific research question rather than a fixed map of the entire area. In addition, the field is developing quickly, so any organization offered here will inevitably remain provisional.

Ethics Statement

This paper surveys prior work and does not introduce a new model or collect new human data. However, selective reduction methods can still create downstream risks. Overly aggressive filtering may suppress minority cases, discard safety-relevant information, or increase confidence on a narrower slice of tasks while weakening broader reliability. We therefore encourage future work to evaluate not only average gains, but also coverage, fairness, and robustness under shift.

AI Assistance Statement

AI-based tools were used to assist with language polishing during the preparation of this manuscript.

References

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. *arXiv preprint arXiv:2402.16827*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Yicheng Chen, Yining Li, Kai Hu, Ma Zerun, HaochenYe HaochenYe, and Kai Chen. 2025. MIG: Automatic data selection for instruction tuning by maximizing information gain in semantic space. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Daewon Choi, Jimin Lee, Jihoon Tack, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. 2025. Think clearly: Improving reasoning via redundant token pruning.

Qirun Dai, Dylan Zhang, Jiaqi W. Ma, and Hao Peng. 2025. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan, Lin, Jan Kautz, and Pavlo Molchanov. 2025. Nemotron-climb: Clustering-based iterative data mixture bootstrapping for language model pre-training. *arXiv preprint arXiv:2504.13161*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning.

Sugyeong Eo, Jungjun Lee, Chanjun Park, and Heuiseok Lim. 2025. Mixture-of-clustered-experts: Advancing expert specialization and generalization in instruction tuning. *arXiv preprint arXiv:2509.10513*.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. 2026. Advancing expert specialization for better moe.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models.

Baizhou Huang, Shuai Lu, Xiaojun Wan, and Nan Duan. 2024. Enhancing large language models in coding through multi-perspective self-consistency. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu.

2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Tingyu Jiang, Shen Li, Yiyao Song, Lan Zhang, Hualei Zhu, Yuan Zhao, Xiaohang Xu, Kenjiro Taura, and Hao Henry Wang. 2025. Importance-aware data selection for efficient llm instruction tuning.

Haoxiang Jin, Ronghan Li, Zixiang Lu, and Qiguang Miao. 2025. Rethinking llm parametric knowledge as post-retrieval confidence for dynamic retrieval and reranking.

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)*. Association for Computational Linguistics.

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

Joshua Ong Jun Leang, Zheng Zhao, Aryo Pradipta Gema, Sohee Yang, Wai-Chung Kwan, Xuanli He, Wenda Li, Pasquale Minervini, Eleonora Giunchiglia, and Shay B. Cohen. 2025. Picsar: Probabilistic confidence selection and ranking for reasoning chains.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Zhuang Li, Yuncheng Hua, Thuy-Trang Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. 2025. 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)*. Association for Computational Linguistics.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2025. Rho-1: Not all tokens are what you need.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2025a. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection.

MingShan Liu and Jialing Fang. 2025. Enhancing mathematical reasoning in large language models with self-consistency-based hallucination detection.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025b. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning.

Quanyu Long, Kai Jie Jiang, Jianda Chen, Xu Guo, Leilei Gan, and Wenya Wang. 2026. Self-verification dilemma: Experience-driven suppression of overused checking in llm reasoning.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. 2025. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16669–16684. Association for Computational Linguistics.

Kyle Montgomery, Sijun Tan, Yuqi Chen, Siyuan Zhuang, Tianjun Zhang, Raluca Ada Popa, and Chenguang Wang. 2025. Budget-aware test-time scaling via discriminative verification.

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*. Association for Computational Linguistics.

Omar Naim, Krish Sharma, Niyar R Barman, and Nicholas Asher. 2026. Tell-tale: Task efficient llms with task aware layer elimination. *arXiv preprint arXiv:2510.22767*.

Nihal V. Nayak, Paula Rodriguez-Diaz, Neha Hulkund, Sara Beery, and David Alvarez-Melis. 2026. A critical look at targeted instruction selection: Disentangling what matters (and what doesn't).

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Zhiwen Ruan, Yixia Li, He Zhu, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. 2025. Enhancing large language model reasoning via selective critical token fine-tuning. *arXiv preprint arXiv:2510.10974*.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Zhipeng Song, Xiangyu Kong, Xinrui Bao, Yizhi Zhou, Jiulong Jiao, Sitong Liu, Yuhang Zhou, and Heng Qi. 2026. Llm-confidence reranker: A training-free approach for enhancing retrieval-augmented generation systems. *Expert Systems with Applications*.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. 2025. Improving pretraining data using perplexity correlations. In *The Thirteenth International Conference on Learning Representations*.

Vernon Y. H. Toh, Deepanway Ghosal, and Soujanya Poria. 2024. Not all votes count! programs as verifiers improve self-consistency of language models for math reasoning.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. 2024. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025. InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too).

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. 2026. Self-routing rag: Binding selective retrieval with knowledge verbalization.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning.

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. *arXiv preprint arXiv:2305.10429*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Jiarui Yao, Ruida Wang, and Tong Zhang. 2025. FANS: Formal answer selection for LLM natural language math reasoning using lean4. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yue Yu, Qiwei Di, Quanquan Gu, and Dongruo Zhou. 2025. On the limits of test-time compute: Sequential reward filtering for better inference.

Shuzhou Yuan, Ercong Nie, Bolei Ma, and Michael Färber. 2025. Why lift so heavy? slimming large language models by cutting off the layers. *arXiv preprint arXiv:2402.11700*.

Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Hui Liu, Yue Xing, Monica Xiao Cheng, and Jiliang Tang. 2025. Towards knowledge checking in retrieval-augmented generation: A representation perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics.

Mingyuan Zhang, Yue Bai, Huan Wang, Yizhou Wang, Qihua Dong, and Yun Fu. 2025a. Boosting large language models with mask fine-tuning. *arXiv preprint arXiv:2503.22764*.

Mingyuan Zhang, Yue Bai, Yifan Wang, Yiyang Huang, and Yun Fu. 2025b. Rethinking fine-tuning: Unlocking hidden capabilities in vision-language models.

Wenxuan Zhang, Paul Janson, Rahaf Aljundi, and Mohamed Elhoseiny. 2024. Overcoming generic knowledge loss with selective parameter update. *arXiv preprint arXiv:2308.12462*.

Shangziqi Zhao, Jiahao Yuan, Jinyang Wu, Zhenglin Wang, Guisong Yang, and Usman Naseem. 2025. Can pruning improve reasoning? revisiting long-cot compression with capability in mind for better reasoning.

Kecheng Zheng, Wei Wu, Ruili Feng, Kai Zhu, Jiawei Liu, Deli Zhao, Zheng-Jun Zha, Wei Chen, and Yujun Shen. 2023. Regularized mask tuning: Uncovering hidden knowledge in pre-trained vision-language models.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini 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.

Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2025. DavIR: Data selection via implicit reward for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Hanna Zubkova, Ji-Hoon Park, and Seong-Whan Lee. 2025. Sugar: Leveraging contextual confidence for smarter retrieval.

Appendix

A More Discussion

The evaluation risks of selective reduction differ across the three stages discussed in this survey. Input-side selection is most vulnerable to coverage loss and selector overfitting: filtering training data or contextual evidence may improve average performance while silently removing rare skills, minority cases, or decisive information. Model-internal selection is more vulnerable to structure mismatch and underfitting: sparse updates, masks, or reduced architectures may preserve useful pre-trained knowledge, but can fail when the selected structure is too restrictive for the downstream task. Inference-time selection is particularly sensitive to false filtering and budget confounding: retrieval filters and reasoning verifiers may improve final answers, but gains can depend heavily on additional scoring compute or on selectors that mistakenly reject useful evidence or valid reasoning paths. Consequently, selective reduction methods should be evaluated not only by their improvements over full baselines, but also by whether they preserve coverage, remain robust under imperfect selection, and generalize across models, budgets, and tasks.

Coverage loss. Selective reduction can introduce hidden blind spots. A reduced dataset, retrieval set, or reasoning pool may outperform the full version on average benchmarks while silently discarding rare capabilities or long-tail cases. This issue is especially important in instruction tuning and retrieval-augmented generation, where average performance can obscure failures on infrequent but important examples.

Selector overfitting. Many methods rely on proxy signals such as uncertainty, confidence, information gain, or stylistic consistency. If these signals become overly tied to specific benchmarks or evaluation settings, selectors may optimize toward the evaluation itself rather than reveal a generally useful reduction principle. This concern is particularly visible in targeted instruction selection, where controlled studies show that reported gains can vary substantially with representation choice, selection strategy, and data budget (Nayak et al., 2026).

False filtering. Knowledge filtering and reasoning selection are inherently vulnerable to false negatives. A retrieved document may appear noisy

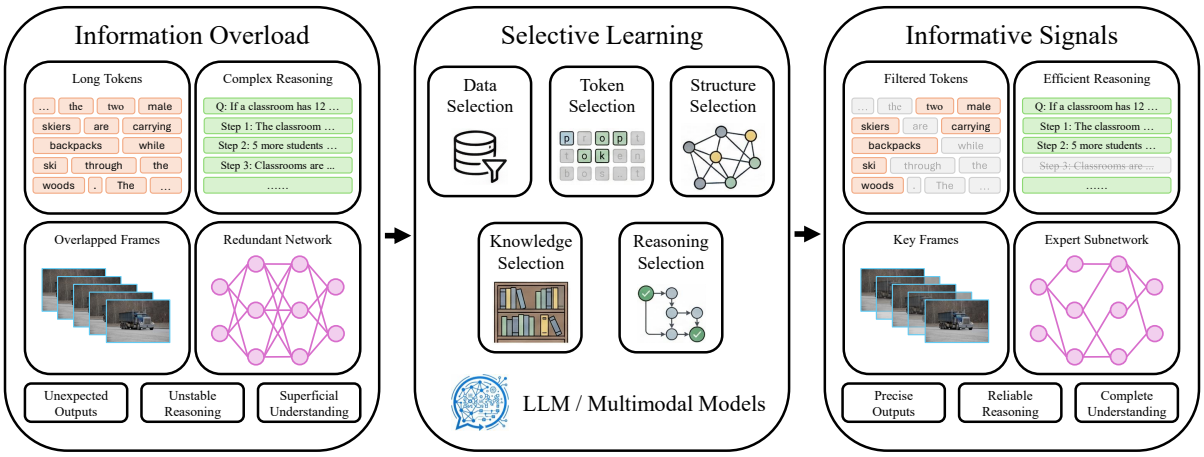

Figure 2: Excessive information, such as input evidence, reasoning candidates, visual frames, and model structure, can introduce noise that degrades model performance. Selective learning identifies and preserves informative components while filtering redundant signals across multiple dimensions. This increases signal density, leading to stronger downstream performance across language and multimodal tasks.

while containing the key supporting fact, and an atypical reasoning trajectory may still lead to the correct answer. Overaggressive filtering can therefore degrade performance even when average-case results improve.

Unfair baselines. The strongest evidence for performance improvements from subtraction comes from comparisons against carefully optimized full-information systems. Weak baselines can make selective methods appear more effective than they truly are. Future work should therefore evaluate whether selective reduction remains beneficial when compared against strong full-data, full-context, and full-reasoning baselines rather than naive configurations alone.

Budget sensitivity. Many reported gains are highly dependent on computational budget. In some cases, selective systems implicitly rely on additional compute from rerankers, verifiers, or confidence estimators. Evaluations should therefore report budget-performance trade-offs and distinguish gains arising from better selection from gains introduced by additional scoring capacity.

Mechanistic clarity. To support the broader claims of selective reduction, evaluations should demonstrate not only that a reduced system performs better but also why it does so. Ideally, experiments should provide evidence that selective reduction improves model behavior by changing what the model attends to, retrieves, updates, or verifies. More generally, evaluation protocols should measure not only average accuracy, but also calibration, robustness, and stability across conditions.

Base-model dependence. The effectiveness of selective reduction is often strongly tied to the underlying model family, scale, training recipe, and evaluation setting. A selection strategy that improves one foundation model may not transfer directly to another because different models vary in their internal knowledge, robustness to noisy context, sensitivity to distribution shift, and ability to recover from removed information. Cross-model evaluation is therefore important for distinguishing generally useful selection principles from model-specific effects. Accordingly, the gains summarized in this survey should be interpreted within the settings reported in the original papers rather than as universal estimates of improvement.