Less Can Be Better: A Survey of Performance-Oriented Selection in Large Language and Foundation Models

Anonymous ACL submission

Abstract

Scaling has been the dominant recipe behind recent progress in large language models and related foundation models. Yet a growing body of work shows that stronger performance does not always come from adding more data, more parameters, more context, more retrieved evidence, or more reasoning. In many settings, models improve after a selective reduction step that removes redundant, noisy, misleading, or otherwise low-value information. This survey reviews the literature through a performance-first lens. We focus on methods that *doing less can directly yield better outcomes*. We organize the field into three stages of the model pipeline: input-side selection, model-internal selection, and inference-time selection. Across these stages, we show that diverse techniques share two common mechanisms: they densify useful signal and reduce interference from irrelevant components. We also identify boundary cases, evaluation pitfalls, and open problems for turning selection from a collection of local heuristics into a general design principle for future LLM systems.

1 Introduction

Recent progress in large language models (LLMs) and broader foundation models has been closely associated with scale (Kaplan et al., 2020; Hoffmann et al., 2022). Larger pretraining corpora, longer contexts, more parameters, more retrieved evidence, and more inference-time reasoning have all been used to improve performance (Wei et al., 2023). However, the empirical picture is no longer monotonic. In many practical settings, models are exposed to large amounts of information that are redundant, weakly relevant, or actively distracting. Under these conditions, adding more does not merely saturate returns but degrade the final prediction (Liu et al., 2024a; Shi et al., 2023).

This observation has motivated a rapidly growing literature on selective reduction. Some methods improve alignment by fine-tuning on a small subset of carefully chosen training samples rather than the full dataset (Zhou et al., 2023; Liu et al., 2024b, 2025). Others improve long-context or retrieval-augmented generation by filtering input evidence before it reaches the generator (Jiang et al., 2024; Wang et al., 2023b; Yan et al., 2024). Another group of methods improves downstream behavior by updating only selected model parameters or model structure (Sharma et al., 2023; Zhang et al., 2025a; Naim et al., 2026). At inference time, better outputs can also emerge by keeping only high-quality retrieved evidence and reasoning candidates (Asai et al., 2024; Wang et al., 2023a; Toh et al., 2024).

These methods are usually discussed in separate research threads such as data curation, prompt compression, sparse adaptation, RAG filtering, or test-time reasoning. This makes it hard to see the common principle that unifies them. 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 broader 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*.

We therefore organize the literature around *where* the subtraction occurs. Input-side selection decides which supervision and input evidence should enter the model. Model-internal selection decides which parameters and structures should participate. Inference-time selection determines which retrieved evidence and reasoning candidates should survive to the final prediction. Throughout the survey, we focus on papers in which the reduction step itself improves accuracy, alignment, robustness, or reasoning quality, and we use this framing to compare methods that are typically studied in separate lines of work.

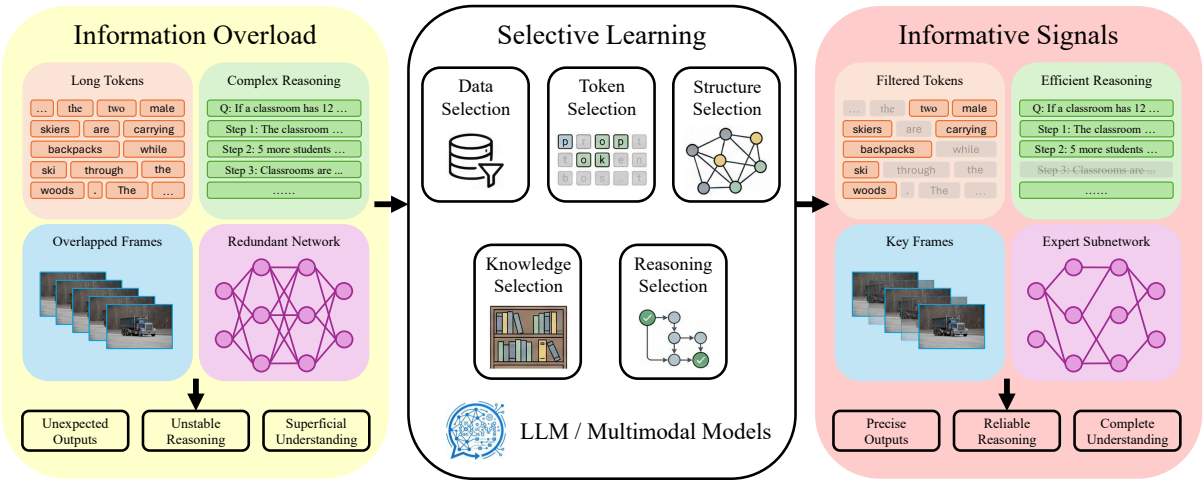

Figure 1: Excessive information, such as input evidence, reasoning candidates, visual frames, and model structure, can introduce noise that degrades model performance. Selective learning identifies and preserves informative components while filtering redundant signals across multiple dimensions. This increases signal density, leading to stronger downstream performance across language and multimodal tasks.

2 Scope and Inclusion Criteria

This survey studies performance-improving selection in LLMs and related foundation models. Our scope centers on cases in which removing, filtering, ranking, masking, routing, or locally updating information improves model behavior.

Inclusion rule. We include work in which selection, filtering, deletion, ranking, masking, routing, or localized updating is a core mechanism, and where the main empirical claim is that this reduction *improves model quality*. Quality may be measured by stronger instruction-following, better reasoning accuracy, better factuality, greater robustness to noisy context, or more balanced capability learning.

Performance criterion. We prioritize papers whose main empirical contribution is that a selective reduction step improves end-task quality. The key outcomes in this survey are stronger instruction-following, better reasoning accuracy, better factuality, greater robustness to noisy context, and better preservation of useful capabilities.

Exclusion rule. We do not aim to cover every form of compression, sparsity, or routing. We exclude work whose main contribution is efficiency alone without evidence of improved task quality over a well-matched full baseline, as well as cases where selection is only an implementation detail rather than the object of study.

Core research question. The central question of the survey is simple: *when does removing something make the model better than before?* This framing keeps the discussion centered on performance-improving subtraction as a modeling principle, rather than as an isolated trick tied to any single mechanism or stage of the pipeline.

Modality scope. The survey is centered on LLMs, but we retain a small number of multimodal or vision-language papers when they provide clean evidence that the same subtraction principle extends beyond text. We use such papers as supporting evidence rather than the main narrative.

3 Why Less Can Be Better

The surveyed methods differ in mechanism, timing, and granularity, but many can be interpreted through two shared ideas.

Signal densification. The first idea is that model quality depends on the density of useful signal, not just the amount of available information. In instruction tuning, weak or low-value examples dilute the update signal; this is exactly why small, curated sets such as LIMA or automatically selected sets such as DEITA and SelectIT can compete with or outperform much larger corpora (Zhou et al., 2023; Liu et al., 2024b, 2025). The same logic reappears in context filtering and prompt compression. If only a fraction of the prompt actually matters, deleting the rest can make the remaining evidence easier for the model to use (Jiang et al., 2024; Weston and Sukhbaatar, 2023).

Interference reduction. The second idea is that irrelevant information is not merely neutral; it

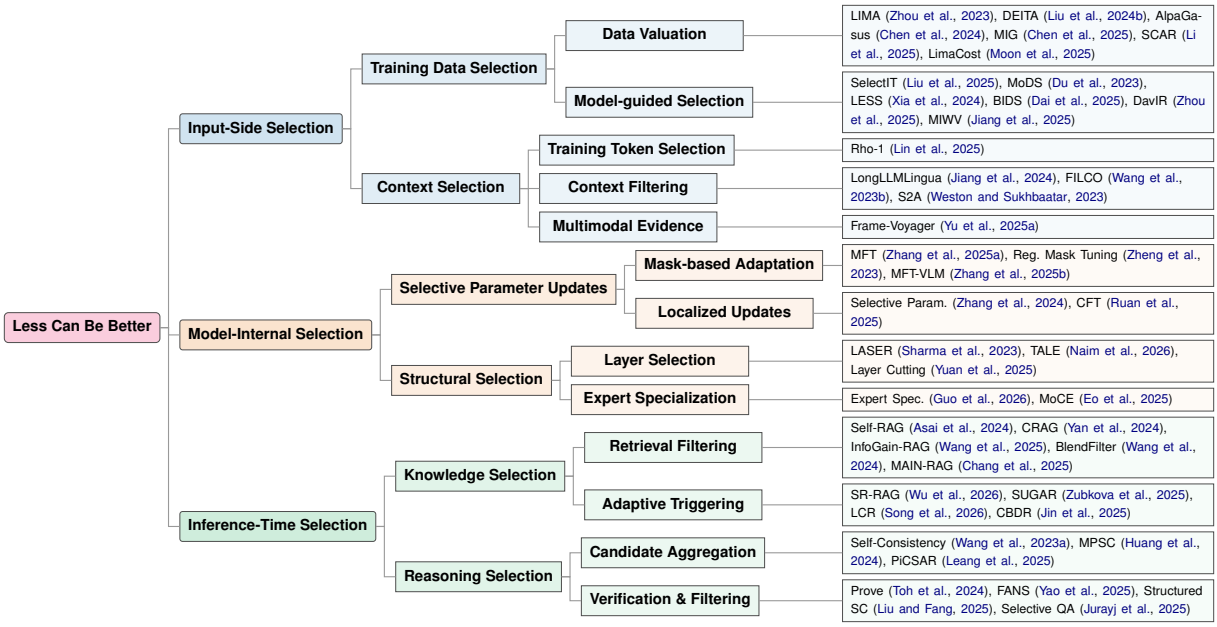

Figure 2: Taxonomy of performance-oriented selection. We organize methods by *where* the subtraction occurs: input-side selection removes low-value training samples or input evidence; model-internal selection activates or updates only task-relevant model parameters or model structure; inference-time selection filters retrieved evidence or reasoning candidates. In each leaf, we list representative methods reviewed in this survey.

can interfere with prediction. Retrieved evidence may conflict with parametric knowledge, weak reasoning candidates may dominate majority voting, and broad parameter updates may damage general knowledge while fitting a narrow task. Methods such as CRAG, Self-RAG, selective parameter update, and reasoning-path verification all succeed partly because they suppress components that would otherwise compete with better evidence (Yan et al., 2024; Asai et al., 2024; Zhang et al., 2024; Toh et al., 2024).

Selection as a performance principle. Taken together, these ideas suggest that subtraction is often a way to reshape the model's information flow, making useful evidence easier to isolate and harmful evidence less likely to dominate. This perspective lets us compare training-sample selection, context selection, structural selection, retrieval filtering, and answer verification within one performance-centered framework.

When less fails. The same perspective also clarifies the limits of the thesis. Reduction is least likely to help when useful evidence is already sparse, when rare cases matter more than average behavior, or when the selector cannot reliably distinguish noise from signal. In short, "less is better" is a conditional claim about redundancy and interference, not a universal preference for smaller inputs or smaller models.

4 Input-Side Selection

Input-side selection concerns what enters the model before or during learning. The relevant question is how to expose the model to *better* supervision and *better* input evidence.

4.1 Training Data Selection

The clearest evidence for the "less is better" hypothesis appears in instruction tuning and alignment. LIMA showed that a 65B LLaMA model fine-tuned on only 1,000 carefully curated examples can display surprisingly strong alignment behavior, suggesting that much of the needed knowledge is already present and that post-training benefits from well-chosen demonstrations rather than sheer volume (Zhou et al., 2023). DEITA sharpened this result by ranking instruction-response pairs according to complexity, quality, and diversity, showing that a much smaller set can retain or improve alignment quality (Liu et al., 2024b). AlpaGasus likewise demonstrated that filtering a noisy dataset to a much smaller, high-quality subset can outperform the original full dataset (Chen et al., 2024).

Subsequent work shifted from generic quality scoring to more model-aware valuation. SelectIT uses uncertainty-aware self-reflection to identify examples that are likely to help the target model

most (Liu et al., 2025). MoDS stages selection around quality, coverage, and necessity, explicitly asking which instructions remain needed after an initial tuning pass (Du et al., 2023). LESS estimates example influence through low-rank gradient similarity and is especially notable because it frames data selection as *task-targeted subtraction*: the value of an example depends on whether it induces the right update for a desired downstream capability (Xia et al., 2024). Related methods such as BIDS and DavIR continue this shift toward model-dependent utility by balancing influence across capability groups or by measuring how learnable and reward-aligned a datum is for the downstream objective (Dai et al., 2025; Zhou et al., 2025). At the same time, recent critical analysis shows that targeted instruction selection is not uniformly beneficial: gains depend strongly on the underlying data representation, the selection algorithm, and especially the budget, with advantages often concentrated at low-budget regimes and diminishing as more data is retained (Nayak et al., 2026). This is useful for our survey because it sharpens the claim from "selection helps" to the more precise claim that *the right kind of subtraction helps under identifiable conditions*.

Another important subline asks what property of the dataset should be preserved when it is reduced. MIG emphasizes information gain and global semantic coverage rather than local heuristics alone (Chen et al., 2025). LimaCost tries to estimate training value relative to the gradient contribution of a strong reference set (Moon et al., 2025). SCAR shows that stylistic consistency can matter even among high-quality examples, because coherent response style may make the supervision signal easier for the model to internalize (Li et al., 2025). Importance-aware data selection reaches a similar conclusion from another angle: for a specific model, the most useful examples may be the ones that expose current weaknesses rather than those that look globally "clean" in the abstract (Jiang et al., 2025).

Across these papers, the recurring lesson is that full data is often not the right target. Performance improves when the selected subset preserves what the model still needs to learn while removing examples that are redundant, low-quality, misaligned, or poorly matched to the current model state.

4.2 Context Selection

Selective reduction also occurs below the sample level. Instead of discarding entire examples, these methods decide which tokens or pieces of evidence should survive.

During training, Rho-1 demonstrates that not all pretraining tokens are equally valuable. By applying loss selectively to useful tokens, it improves mathematical and general downstream performance while using only a fraction of the original token stream (Lin et al., 2025). The result is important because it shows that the "less is better" effect is not limited to choosing whole examples; it can emerge from choosing which micro-signals within each example actually matter.

At inference time, long-context and RAG settings provide especially strong evidence that more tokens can hurt. LongLLMLingua improves long-context performance by compressing prompts in a question-aware way, effectively keeping tokens that matter while removing distractors (Jiang et al., 2024). FILCO and System 2 Attention similarly improve results by filtering or regenerating context so that irrelevant passages exert less influence on the final answer (Wang et al., 2023b; Weston and Sukhbaatar, 2023). Together, these papers show that deleting noisy context can directly improve answer quality.

The same logic extends beyond text. Frame-Voyager demonstrates in video-language settings that a model can improve after selecting a small subset of informative frames rather than processing the full sequence (Yu et al., 2025a). This does not change the survey's center of gravity, which remains LLMs, but it reinforces the general point that selective retention of evidence can be performance-improving in broader foundation-model pipelines. Across these settings, the shared intuition is the same even though the bottleneck differs: token selection reallocates the learning signal, context filtering reshapes attention at test time, and multimodal selection reduces cross-modal redundancy.

5 Model-Internal Selection

Model-internal selection asks whether better performance can arise by activating, updating, or preserving only part of the model. We focus on papers in which reducing model parameters or model structure is directly argued to improve the task outcome.

5.1 Selective Parameter Updates

One major line of work starts from the hypothesis that much of the needed capability already exists inside pretrained weights, but standard full updates disturb it too broadly. Mask Fine-Tuning (MFT) makes this claim explicit by learning masks over existing parameters, rather than treating all weights as equally appropriate for adaptation (Zhang et al., 2025a). Regularized Mask Tuning reaches a similar conclusion in vision-language transfer, arguing that useful downstream behavior can emerge by uncovering relevant pathways rather than by indiscriminately rewriting the full network (Zheng et al., 2023). The same intuition also appears in broader vision-language evidence, where mask-based adaptation can outperform standard low-rank updating by more effectively reusing the structure already present in pretrained weights (Zhang et al., 2025b).

Selective parameter update pushes this logic further by restricting updates to a sparse task-relevant subset so that new knowledge is added without excessive damage to generic knowledge (Zhang et al., 2024). Critical Token Fine-tuning offers a more localized variant for reasoning tasks: instead of backpropagating equally through every token, it updates the model around counterfactually identified critical reasoning tokens and reports stronger mathematical reasoning despite touching fewer supervised positions (Ruan et al., 2025). In both cases, the core idea is that over-broad adaptation introduces harmful interference, whereas localized adaptation preserves useful general structure.

5.2 Structural Selection

A second line of work performs subtraction at the level of larger internal structures. LASER shows that selectively reducing rank in particular layers can improve reasoning after training is already complete, implying that some higher-order components are not merely unnecessary but detrimental for certain tasks (Sharma et al., 2023). TALE makes a similar point at inference time by removing task-irrelevant layers and reporting that the resulting task-adapted architecture can match or surpass the original full model (Naim et al., 2026). Even more provocatively, work on cutting off layers reports that fewer layers can sometimes outperform the intact model for particular prompt-based tasks (Yuan et al., 2025).

Mixture-of-Experts (MoE) papers enter this survey when the main argument is that *better selection among experts improves prediction quality*. Under that criterion, work on expert specialization is relevant. Encouraging more distinct experts can improve task performance because it reduces overlap and prevents routing from collapsing toward mediocre average behavior (Guo et al., 2026). Mixture-of-Clustered-Experts extends this idea by partitioning inputs more effectively before fine-grained expert routing and reporting gains from better specialization (Eo et al., 2025). Unlike selective parameter updates, which ask which weights should change, structural selection asks which layers or experts should participate at all.

Overall, model-internal selection supports a strong version of the survey's thesis: even after the model has already been trained, better performance can come from *not* using the entire network uniformly. Selective updating and selective activation matter because capabilities are unevenly distributed, and broad participation can introduce destructive interference.

6 Inference-Time Selection

Inference-time selection asks which retrieved evidence and reasoning candidates should survive until the final answer. The underlying model is fixed; what changes is the policy for choosing what the model should rely on.

6.1 Knowledge Selection

RAG systems provide a particularly clear example of why more information can hurt. Retrieved evidence often contains irrelevant, conflicting, or low-value material. The practical challenge is therefore not only to retrieve more, but to decide whether retrieval is needed and which retrieved evidence should be kept.

Self-RAG internalizes part of this decision by teaching the model to retrieve, critique, and use evidence selectively rather than consuming a fixed number of passages every time (Asai et al., 2024). CRAG adds an explicit retrieval evaluator and corrective pathway, showing that quality improves when the system detects poor retrieval and filters or repairs it before generation (Yan et al., 2024). InfoGain-RAG sharpens the idea further by scoring retrieved documents according to how much they improve answer generation, converting document selection into a direct estimate of utility to the generator (Wang et al., 2025). Related methods such as BlendFilter, MAIN-RAG, representation-

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

based knowledge checking, and DRAGIN differ in mechanism, but converge on the same claim that RAG improves when the system filters knowledge according to usefulness rather than letting all retrieved material compete equally (Wang et al., 2024; Chang et al., 2025; Zeng et al., 2025; Su et al., 2024).

Another branch of work makes the selection decision earlier, before or during retrieval itself. Self-Routing RAG treats selective retrieval as knowledge-source selection, allowing the model to decide whether to retrieve externally or rely on parametric knowledge (Wu et al., 2026). SUGAR similarly uses contextual confidence to trigger retrieval only when it seems helpful (Zubkova et al., 2025). LLM-confidence reranking and post-retrieval confidence estimation show that even after evidence is retrieved, the model's own confidence signals can help identify which evidence is worth trusting (Song et al., 2026; Jin et al., 2025). The common pattern is simple but powerful: less retrieved evidence, if it is better chosen, can produce more accurate and more stable answers than a larger but noisier retrieval set.

6.2 Reasoning Selection

Selection can also happen after the model has already generated reasoning candidates. Self-Consistency is the canonical example: instead of trusting a single greedy chain of thought, it samples multiple reasoning candidates and selects the answer supported most consistently across them (Wang et al., 2023a). Multi-Perspective Self-Consistency extends this idea by comparing candidates across multiple generated views, such as solutions, specifications, and tests, thereby improving quality through selection among alternatives rather than relying on a larger base model (Huang et al., 2024). PiCSAR shows that even within best-of-N style reasoning, better scoring of candidate chains can reduce the number of samples needed while improving final selection quality (Leang et al., 2025). More recent work makes the same point under explicit computational constraints. Sequential reward filtering improves over standard best-of-N style inference by selectively carrying only high-reward generations forward, concentrating compute on promising candidates instead of letting weaker trajectories remain in the loop (Yu et al., 2025b). Budget-aware test-time scaling via discriminative verification reaches a similar conclusion: under a fixed budget, selective verification can outperform more expensive generative verification, suggesting that better filtering can matter more than simply searching harder (Montgomery et al., 2025).

Verification-based methods make the subtraction even more explicit. Programs-as-Verifiers filters candidate math solutions by checking whether translated programs agree with the natural-

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)

language reasoning (Toh et al., 2024). FANS moves to formal verification with Lean4, selecting only those answers that survive stronger symbolic constraints (Yao et al., 2025). Structured self-consistency for mathematics similarly reports that gains come not only from choosing the most common final answer, but from discarding intermediate reasoning that is internally inconsistent (Liu and Fang, 2025). A complementary line asks whether parts of the reasoning trace itself should be removed. Think Clearly reports that pruning redundant reasoning tokens can improve downstream accuracy, especially when long reasoning trajectories become diffuse and distracting (Choi et al., 2025). Related evidence from long-CoT compression shows that not all pruning is equal: removing low-value verification segments can help, whereas indiscriminate pruning of core reasoning steps can degrade performance, making the benefit depend on *which* parts of the chain are suppressed (Zhao et al., 2025). In a similar spirit, Self-Verification Dilemma shows that reflective checking is often overused and mostly confirmatory; suppressing unnecessary rechecks can preserve or even improve accuracy while reducing token use (Long et al., 2026). In selective question answering, abstention or confidence-thresholding adds another angle: sometimes the best way to improve decision quality is to answer fewer questions, but answer them more reliably (Jurayj et al., 2025).

This part of the literature clarifies an important nuance. Inference-time improvement does not always require *more* search. What often matters is which candidates are retained, which are verified, and which are rejected before they contaminate the final decision.

7 Evaluation Principles

The literature reviewed here is compelling, but it is also easy to overclaim. A method can look strong under selective reduction for reasons that do not really support the "less is better" thesis.

Coverage loss. Removing data, layers, or documents can create brittle blind spots. A selected subset may outperform the full set on a narrow benchmark while silently dropping rare capabili-

ties or minority cases. This is especially important for instruction tuning, where strong average scores can hide poor coverage of long-tail skills.

Selector overfitting. Many methods depend on proxy scores such as uncertainty, confidence, information gain, or style consistency. If the selector is too closely tied to one benchmark family, it may overfit the evaluation rather than reveal a general performance principle. This concern is especially visible in targeted instruction selection, where recent controlled comparisons show that apparent gains can vary substantially with representation choice, selection algorithm, and budget (Nayak et al., 2026).

False filtering. Knowledge selection and reasoning filtering are vulnerable to false negatives. A document may look noisy but contain the decisive fact; a reasoning chain may look atypical but still be correct. Overaggressive filtering can therefore hurt even when the average case improves.

Unfair baselines. The strongest evidence for performance-improving subtraction comes from comparisons against well-trained full-data, full-context, or full-reasoning baselines. Weak baselines can make selection look better than it is. Future work should report whether gains survive against carefully tuned full-information systems rather than naive baselines alone.

Budget sensitivity. Many gains are highly budget-dependent, and some selective systems also borrow substantial extra compute from selectors, verifiers, or rerankers. Strong evaluations should therefore report budget-performance curves and separate gains from better retained information from gains due to additional scoring capacity.

Mechanistic clarity. To support the argument of this survey, evaluations should show that the selected configuration helps by changing *what the model attends to or updates*. More broadly, evaluation should be matched to the stage at which selection operates and should measure not only average quality, but also calibration, and robustness.

8 Open Problems

One open problem is how to coordinate selection across stages. Current methods typically optimize a single local decision at a time, such as selecting training samples, filtering documents, or pruning reasoning paths. A more general theory would ask how these decisions interact across the pipeline. For example, a model trained on selectively curated data may require a different retrieval policy at inference time than a model trained on the full distribution.

Another open problem is how to move from heuristic selectors to end-task-grounded selection. Many current methods rely on plausible proxies, but the field still lacks a unified account of when uncertainty, influence, confidence, or information gain should be trusted as signals for reduction. Better causal or optimization-based analyses could make the principle more predictable and less dataset-specific.

A third challenge concerns robustness under distribution shift. Selective systems often work because they learn what is usually irrelevant. But under shift, previously discardable signals may become decisive. This is especially important in RAG, multimodal evidence selection, and expert routing, where the useful subset can change rapidly with the query or task.

Finally, the field needs clearer benchmarks for performance-oriented subtraction. A mature evaluation should ask not only whether a selector beats a full baseline, but whether it does so while preserving coverage, maintaining calibration, and remaining stable across task families. More broadly, the field still lacks principled joint selection across stages and stronger counterfactual evidence for why many proposed selection signals work.

9 Conclusion

This survey examined a simple but increasingly important idea: stronger model performance does not always require more. Across input curation, internal adaptation, and inference-time decision making, a wide range of recent work suggests that selective reduction can improve quality by densifying useful signal and reducing interference from distracting components. Framed this way, "less is better" is not an isolated slogan from alignment or prompt compression. It is a broader design principle for LLM systems and related foundation models. At the same time, the evidence surveyed here also shows that the principle is conditional rather than universal: reduction helps most when redundancy and interference are substantial, and it can fail when coverage, rare evidence, or selector reliability become the dominant concern. As models continue to grow, progress may depend not only on how much information and capacity they can absorb, but on how effectively they learn what to leave out.

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 new human data collection. 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 and literature organization during the preparation of this manuscript.

References

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.

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. *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.

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. 2025. 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*.

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.

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. *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).

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. *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*.

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.

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.

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.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. 2025a. Frame-voyager: Learning to query frames for video large language models.

Yue Yu, Qiwei Di, Quanquan Gu, and Dongruo Zhou. 2025b. 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. *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. *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. *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.