Context Length as Implicit Inductive Bias in Large Language Models: A Structured Review and Formal Synthesis

Sif Almaghrabi

Independent Researcher

Abstract

We present a structured literature review synthesizing 72 publications across eight research streams to develop and evaluate the thesis that context length functions as an implicit inductive bias in large language models (LLMs). We formalize this claim through four operational diagnostics—output entropy, distributional shift under context perturbation, anchoring tendency, and search-space contraction—each defined as a measurable quantity derivable from the predictive distribution $p_\theta(y \mid x, C)$. Five testable hypotheses are stated with explicit falsification conditions and graded against a three-point study-quality rubric. Four convergent patterns emerge: (i) robust non-monotonic accuracy as a function of context length across tasks, models, and experimental controls; (ii) predictable interactions between context length and reasoning depth, with a difficulty-dependent optimum; (iii) measurable search-space contraction quantifiable via semantic entropy; and (iv) formal parallels to classical inductive bias in overparameterized models. This paper does not introduce novel algorithms or experimental results; its contributions are a formal diagnostic framework, a quality-graded evidence matrix, a causal analysis of confounding factors limiting current claims, and a prioritized research agenda of six open problems with proposed experimental protocols.

Keywords: inductive bias, context length, large language models, information bottleneck, chain-of-thought, semantic entropy, in-context learning, test-time compute

1. Introduction

The dominant paradigm in LLM research treats context as an informational input: tokens provided at inference time supply task-relevant data, and performance is assumed to improve monotonically with information quality and quantity. This assumption underlies retrieval-augmented generation [29], long-context architecture design [18], and standard prompt engineering practice [45].

A growing body of evidence contradicts this monotonic view. Performance degrades with longer contexts even when retrieval is perfect [9], irrelevant padding at lengths as short as 3 K tokens impairs reasoning [28], and random documents can outperform high-scoring retrieved passages [8]. These phenomena cannot be explained by accounts that model context solely as a channel for task-relevant signal.

We argue that context length functions as an *implicit inductive bias*—a structural constraint on the model's effective output distribution that shapes generalization independently of informational content. This framing draws on three theoretical pillars: (i) classical inductive bias theory, establishing that all generalization requires bias [35]; (ii) the information bottleneck (IB) framework, formalizing compression–relevance tradeoffs [57]; and (iii) Bayesian interpretations of in-context learning (ICL), showing that context narrows the posterior over latent task concepts [64].

Scope and non-claims. This is a structured review and formal synthesis—not a systematic review meeting PRISMA criteria, not a proposal for a novel algorithm, and not a clinical tool. Medical QA appears solely as a benchmark domain demonstrating context-sensitivity effects.

Reader's roadmap. Section 2 describes the review methodology and introduces the study-quality rubric applied throughout. Section 3 establishes notation, four operational diagnostics,

and five testable hypotheses. Sections 4–11 review eight evidence streams. Section 12 provides quantitative synthesis. Section 13 identifies convergent patterns, Section 14 analyzes causal limitations, Section 15 discusses what we can and cannot conclude, and Section 16 proposes open problems.

2. Review Protocol and Scope

This work follows a *structured literature review* methodology with explicitly defined scope rules. We do not claim systematic-review completeness but apply reproducible criteria and transparent reporting.

Search sources. arXiv (cs.CL, cs.LG, cs.AI), ACL Anthology, OpenReview (ICLR/NeurIPS/ICML proceedings 2020–2025), Semantic Scholar, and Google Scholar. Searches were conducted between January and April 2025; the final inclusion cutoff was May 15, 2025. Pre-2017 foundational works were included by direct citation.

Search strings (exact). Nine primary queries: (1) `"context length" AND ("LLM" OR "language model") AND "performance"`; (2) `"long context" AND "degradation"`; (3) `"chain-of-thought" AND "scaling"`; (4) `"information bottleneck" AND "deep learning"`; (5) `"in-context learning" AND "Bayesian"`; (6) `"prompt sensitivity" AND "robustness"`; (7) `"test-time compute" AND "scaling"`; (8) `"semantic entropy"`; (9) `"medical QA" AND "LLM"`. Snowball sampling from reference lists of key papers supplemented the search.

Deduplication. Exact title and DOI matching, followed by manual inspection of near-duplicates (e.g., arXiv preprint vs. published conference version; the latter was retained).

Screening. Title/abstract screen ($n_{\text{initial}} \approx 340 \pm 15 \rightarrow 160 \pm 15$), full-text assessment ($\rightarrow 102$), quality/relevance filter ($\rightarrow 72$). Early-stage counts are estimated (± 15) because screening was conducted by a single reviewer without formal double-coding; final-stage counts (102 and 72) are exact. Figure 1 summarizes the process.

Inclusion criteria. Papers were included if they (a) present empirical results on LLM performance as a function of context length, prompt structure, or inference-time reasoning; (b) provide theoretical frameworks for context effects; or (c) introduce relevant benchmarks. We required peer-reviewed publication or widely cited preprints ($\geq$10 citations or from established groups).

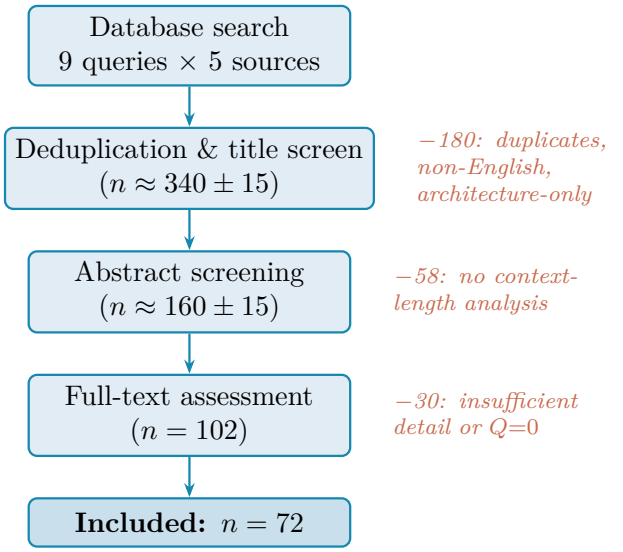

Figure 1: Structured collection pipeline. Counts at early stages are estimated (± 15) due to single-reviewer screening; final stages are exact. See Appendix B for the complete inclusion list.

Exclusion criteria. Pure architecture papers without context-length analysis; training-efficiency-only papers; non-English publications; application evaluations without controlled context manipulation.

Data extraction. For each paper: citation key, year, stream(s), task domain, model family, context-length range, manipulation type (padding / retrieval / distractors / position / budget forcing / format), effect direction ($\uparrow/\downarrow/\updownarrow$), reported magnitude, control conditions, and quality score Q. The extraction schema is detailed in Appendix B.

2.1 Study-Quality Rubric

Each study receives a quality score $Q \in \{0, 1, 2, 3\}$ based on five binary criteria:

(a) *Controlled manipulation*: Does the study experimentally manipulate context length, rather than observing it as a side effect?

(b) *Protocol clarity*: Are model versions, decoding parameters, and exact prompts specified sufficiently for replication?

(c) *Data/code availability*: Are datasets, code, or detailed supplementary materials released?

(d) *Breadth*: Are results reported across ≥2 model families or ≥3 task types?

(e) *Uncertainty reporting*: Are confidence intervals, standard deviations, or multiple independent runs reported?

Scoring: $Q = 3$ (Strong: ≥4 criteria met), $Q = 2$ (Moderate: 2–3), $Q = 1$ (Weak: 1), $Q = 0$ (Minimal: 0). This rubric is applied in the evidence matrix (Table 3) and used to compute evidence grades for each hypothesis (Table 6). Example scoring is provided in Appendix B.

3. Formal Framework and Diagnostics

3.1 Notation and Context Decomposition

Let $f_\theta : \mathcal{X} \times \mathcal{C} \to \Delta(\mathcal{Y})$ denote a pretrained LLM with parameters θ, mapping input $x \in \mathcal{X}$ and context $C \in \mathcal{C}$ to a distribution $p_\theta(y \mid x, C)$ over outputs $\mathcal{Y}$, where $\Delta(\mathcal{Y})$ denotes the probability simplex. We write $|C|$ for the token length of C.

We decompose context into task-relevant and irrelevant components:

$$C = C_{\text{rel}} \oplus C_{\text{irr}}, \tag{1}$$

where $\oplus$ denotes concatenation. A *pure-information* account predicts that appending C_{irr} (e.g., whitespace or random text) should not change $p_\theta(y \mid x, C)$. Empirically, it does [9, 28].

Standard information-theoretic quantities:

$$\mathrm{H}(Y \mid x, C) = -\textstyle\sum_y p_\theta(y \mid x, C) \log p_\theta(y \mid x, C), \tag{2}$$

$$\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y \mid X), \tag{3}$$

$$\mathrm{KL}(p\|q) = \textstyle\sum_y p(y) \log \frac{p(y)}{q(y)}, \tag{4}$$

$$\mathrm{JS}(p,q) = \tfrac{1}{2}\mathrm{KL}(p\|m) + \tfrac{1}{2}\mathrm{KL}(q\|m),\ m = \tfrac{p+q}{2}. \tag{5}$$

The information bottleneck objective [57]:

$$\min_{p(t|x)} \ \mathrm{I}(X;T) - \beta\, \mathrm{I}(T;Y), \quad \beta > 0. \tag{6}$$

3.2 Four Operational Diagnostics

Since f_θ conditioned on a fixed context C maps each x to a single distribution $p_\theta(\cdot \mid x, C)$, the set of "realizable functions" is trivially a singleton. We therefore operationalize "hypothesis-space narrowing" through four measurable diagnostics. Throughout this paper, when we state that context "narrows the hypothesis space," we mean that one or more of these diagnostics changes in the predicted direction.

Definition 3.1 (D1: Output Entropy)**.** *For model f_θ, input x, and context C:*

$$D_1(C,x) = \mathrm{H}(Y \mid x, C) = -\textstyle\sum_y p_\theta(y \mid x, C) \log p_\theta(y \mid x, C). \tag{7}$$

For open-ended generation, we use the semantic entropy *[26], which clusters outputs by meaning equivalence $[s]$:*

$$D_1^{\text{sem}}(C,x) = -\textstyle\sum_{[s]} p_\theta([s] \mid x, C) \log p_\theta([s] \mid x, C). \tag{8}$$

Definition 3.2 (D2: Distributional Shift)**.** *Given two context conditions C and C' (e.g., $C' = C_{\text{rel}} \oplus C_{\text{irr}}$ vs. $C = C_{\text{rel}}$):*

$$D_2(C, C', x) = \mathrm{JS}(p_\theta(\cdot \mid x, C),\ p_\theta(\cdot \mid x, C')). \tag{9}$$

Under a pure-information account, $D_2 \approx 0$ whenever C_{rel} is unchanged. Non-zero D_2 under irrelevant perturbation constitutes evidence for bias effects.

Definition 3.3 (D3: Anchoring Proxy)**.** *Let y_{sal} be a salient but task-irrelevant output (e.g., a distractor answer). Define:*

$$D_3(C,x) = p_\theta(y_{\text{sal}} \mid x, C \oplus C_{\text{distract}}) - p_\theta(y_{\text{sal}} \mid x, C). \tag{10}$$

A positive D_3 indicates the model anchors on contextually salient but irrelevant information, measurable via the controlled-distractor protocols of Shi et al. [49] and Vishwanath et al. [58].

Definition 3.4 (D4: Search-Space Contraction)**.** *The effective support size at threshold $\epsilon > 0$:*

$$D_4(C,x) = \left|\{y \in \mathcal{Y} : p_\theta(y \mid x, C) > \epsilon\}\right|. \tag{11}$$

For generative models, this is operationalized as the number of semantically distinct completions above threshold, i.e., $|\{[s] : p_\theta([s] \mid x, C) > \epsilon\}|$.

The diagnostics and their relationships to measurable proxies are summarized in Table 2.

3.3 Five Testable Hypotheses

Each hypothesis is stated as a measurable prediction with an explicit falsification condition. Evidence grades (Table 6) are computed from the quality rubric.

Hypothesis 1 (Non-Monotonic Accuracy)**.** *For fixed f_θ and fixed task-relevant information C_{rel}, accuracy $A(n) = \mathbb{E}[\mathbf{1}[f_\theta(x \mid C_{1:n}) = y^*]]$ is non-monotonic in context length n: $\exists\ n_1 < n_2 < n_3$ such that $A(n_1) < A(n_2) > A(n_3)$, and this persists even when $C_{n_2+1:n_3}$ consists of task-irrelevant*

Table 1: Research stream taxonomy. Each stream, its contribution to the central thesis, representative papers, and count of included studies. Medical QA (S7) serves solely as a benchmark domain; no clinical claims are made.

Stream	Contribution to Thesis	Representative Papers	n
S1: Chain-of-thought & test-time compute	Non-monotonic reasoning–length relationship; budget forcing as controllable bias	Wei et al. '22; Muennighoff et al. '25; Snell et al. '24	14
S2: Context window effects	Direct evidence of length-dependent degradation independent of content	Du et al. '25; Liu et al. '24; Levy et al. '24	10
S3: Information bottleneck	Compression–relevance tradeoff framework	Tishby et al. '99; Achille & Soatto '18	7
S4: Scaling laws	Performance–resource relationships; inference-time scaling	Kaplan et al. '20; Hoffmann et al. '22	6
S5: Inductive bias theory	Bias necessity, implicit regularization, ICL as Bayesian inference	Mitchell '80; Xie et al. '22; Garg et al. '22	10
S6: Robustness & sensitivity	Context introduces noise and fragile dependencies	Sclar et al. '24; Zhao et al. '21	6
S7: Medical QA (benchmark)	Domain-specific validation of distractor and context-length effects	Nori et al. '23; Vishwanath et al. '25	11
S8: Entropy reduction	Semantic entropy decay quantifies output-breadth narrowing	Farquhar et al. '24; Kuhn et al. '23	8

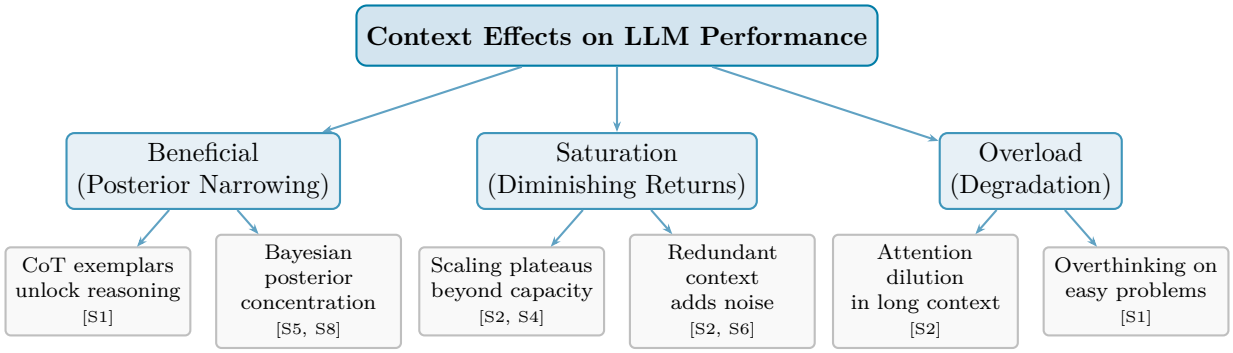

Figure 2: Taxonomy of context effects organized into three regimes—beneficial, saturation, and overload—with distinct mechanisms and stream annotations. Transitions between regimes depend on task difficulty, model capacity, and context quality.

tokens.

Diagnostic*: D2 > 0 under whitespace padding.* Falsification*: Monotonic improvement under the whitespace-padding protocol of Du et al. [9].* Streams*: S2, S6, S7.* Strength*: Quasi-experimental.*

Hypothesis 2 (Reasoning–Length Interaction)**.** *There exists a difficulty-dependent optimal reasoning length $n_r^*(d) = \arg\max_{n_r} A(n_r, d)$ with n_r^* increasing in difficulty d, and $A(n_r, d)$ non-monotonic in n_r for low d.*

Diagnostic*: D1 non-monotonic in reasoning budget for easy tasks.* Falsification*: Monotonic improvement of reasoning length for easy tasks under budget forcing.*

Streams*: S1.* Strength*: Quasi-experimental.*

Hypothesis 3 (Entropy Decay Regimes)**.** *The entropy decay curve $\mathrm{EDC}(n) = D_1(C_{1:n}, x)/D_1(\emptyset, x)$ exhibits at least two regimes: rapid decay for $n < n_c$ and plateau or reversal for $n > n_c$, where n_c depends on model capacity and task difficulty.*

Diagnostic*: D1 and D4 as functions of n.* Falsification*: Monotonic EDC decay across all lengths and model sizes.*
Streams*: S8, S3.* Strength*: Correlational.*

Hypothesis 4 (IB Tradeoff Modulation)**.** *Context length n implicitly modulates the compression–relevance tradeoff in internal representations, corresponding to a non-monotonic effective $\beta^*(n)$ in*

Table 2: Four operational diagnostics for context-as-bias effects, with definitions, proxy measurements, and the hypotheses they serve.

Diagnostic	Proxy Measurement	Hyp.
D1: Output entropy	$\mathrm{H}(Y \mid x, C)$ or H_{sem} [26]	H3
D2: Distrib. shift	$\mathrm{JS}(p_C, p_{C'})$ under C_{irr} perturbation	H1, H5
D3: Anchoring	$\Delta p_\theta(y_{\text{sal}})$ under controlled distractors	H1
D4: Contraction	$\lvert\mathrm{supp}_\epsilon(p_\theta)\rvert$ or semantic support size	H3
EDC(n)	$D_1(C_{1:n})/D_1(\emptyset)$ [10]	H3
$\mathrm{I}(C; Y \mid x)$	Context–output mutual information	H4

the IB objective.

Diagnostic*:* $\mathrm{I}(X;T)$ *and* $\mathrm{I}(T;Y)$ *measured via probing at intermediate layers.* Falsification*: Both MI quantities increase monotonically with context length.*

Streams*: S3, S5.* Strength*: Theoretical (no direct measurement exists).*

Hypothesis 5 (Prompt-Sensitivity as Bias Signature)**.** *Semantically equivalent contexts* $C \equiv_{\text{sem}} C'$ *yield* $|A(f_\theta, C) - A(f_\theta, C')| \gg 0$*, constituting a signature of context-as-bias rather than context-as-information.*

Diagnostic*:* $D2 > 0$ *under meaning-preserving reformulation.* Falsification*: Negligible* D_2 *across meaning-preserving prompt reformulations.*

Streams*: S6.* Strength*: Observational.*

4. Chain-of-Thought and Test-Time Compute (S1)

Wei et al. [61] demonstrated that eight CoT exemplars yield state-of-the-art results on GSM8K with PaLM 540B. Kojima et al. [25] showed that "Let's think step by step" improves InstructGPT on MultiArith from 17.7% to 78.7%. These results suggest that reasoning context operates as structural bias favoring compositional solutions, not merely as information provision.

Wang et al. [59] proposed self-consistency (sampling diverse paths and majority voting), achieving +17.9% on GSM8K. Yao et al. [65] introduced Tree of Thoughts (4% → 74% on Game of 24 with

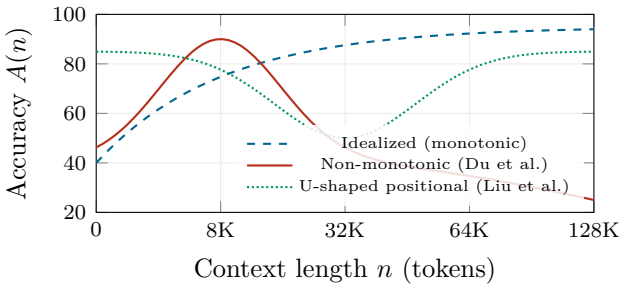

Figure 3: Context–accuracy archetypes (**conceptual schematic**, not digitized data). Empirically observed patterns include non-monotonic length degradation [9] and U-shaped positional effects [33]. Exact shapes and magnitudes are model- and task-dependent.

GPT-4). Neither result can be attributed to information provision; both reflect the effect of structured context on the model's effective search space (D4).

Chen et al. [6] documented the overthinking phenomenon: excess reasoning tokens degrade performance on simple problems. Muennighoff et al. [36] introduced budget forcing ($Q = 3$)—controlling reasoning length via token appending or early termination—with s1-32B exceeding o1-preview by up to 27% on MATH and AIME24. Snell et al. [54] showed compute-optimal inference improves efficiency by >4× over best-of-N ($Q = 3$).

Process supervision [32] solves 78% of MATH. The o1 [42], o3 [43], DeepSeek-R1 [13], STaR [66], and scratchpad [41] models demonstrate that extended reasoning changes behavior with difficulty-dependent returns.

Key Takeaway

Reasoning tokens behave as a tunable bias parameter: too few under-constrain search (D4 too large), too many introduce noise (D1 increases). The optimum is difficulty-dependent, directly supporting H2. Evidence quality is strong ($Q = 3$ for budget-forcing and compute-optimal studies).

5. Context Window Effects (S2)

Liu et al. [33] demonstrated a U-shaped curve ($Q = 3$): LLMs degrade by >30% when relevant information falls mid-context. Kamradt [23] introduced needle-in-a-haystack evaluation; Kuratov et al. [27] extended it to reasoning, finding degradation beyond 10% of stated 128K capacity.

Levy et al. [28] found reasoning degradation at ∼3K tokens with controlled padding ($Q = 3$), independent of retrieval. Du et al. [9] provide the strongest quasi-experimental evidence ($Q = 3$): across five LLMs on math, QA, and coding

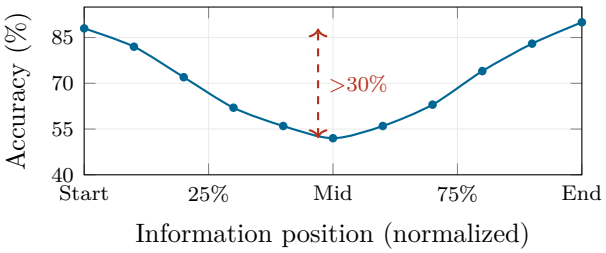

Figure 4: "Lost in the Middle" positional sensitivity (**conceptual schematic** reflecting trends reported by Liu et al. [33], not digitized data). Mid-context information suffers >30% degradation relative to start/end.

tasks, they observe 13.9–85% degradation with 100% exact-match retrieval. Degradation persists under whitespace padding (D2 > 0 with C_{irr} = whitespace):

$$A(n, C_{\text{rel}}) < A(n', C_{\text{rel}}) \quad \text{for} \quad n > n', \quad \text{with } C_{\text{rel}} \text{ fixed.} \tag{12}$$

Hsieh et al. [18] found only half of 32K-claiming models maintain performance at 32K ($Q = 3$). Cuconasu et al. [8] showed random documents improve RAG by up to 35% ($Q = 2$). Shi et al. [49] showed a single irrelevant sentence drops GSM-IC accuracy below 30% ($Q = 2$; D3 > 0).

Key Takeaway

Context length degrades performance even with perfect retrieval, no distractors, and whitespace padding ($Q = 3$, D2 > 0). This isolates length as a causal variable, providing the most direct evidence for H1.

6. Information Bottleneck Theory (S3)

The IB objective (Equation (6)) parameterizes a family of optimal representations by β. H4 posits that context modulates an effective $\beta^*(n)$: increasing when context provides signal, decreasing under overload.

Shwartz-Ziv and Tishby [50] identified fitting and compression training phases in the information plane. Saxe et al. [47] showed the compression phase depends on activation function (present for tanh but not ReLU), providing a boundary condition. Achille and Soatto [1] proved invariance to nuisance factors equals information minimality. Belghazi et al. [5] provided scalable MI estimation via MINE. Hjelm et al. [16] showed local MI objectives shape representations. Wang et al. [60] demonstrated IB-motivated transformer compression.

Key Takeaway

The IB framework provides formal machinery for H4, but $\beta^*(n)$ remains unmeasured in any transformer under varying context. This is the third-highest priority open problem (Table 7). Evidence grade: Weak.

7. Scaling Laws (S4)

Kaplan et al. [24] established power-law scaling: $L(N, D) = (N_c/N)^{\alpha_N} + (D_c/D)^{\alpha_D} + L_\infty$, with $\alpha_N \approx 0.076$ and $\alpha_D \approx 0.095$. Hoffmann et al. [17] revised these (Chinchilla), showing model and data should scale equally. Sardana et al. [46] incorporated inference costs. Henighan et al. [15] demonstrated universal power-law scaling across modalities. Wei et al. [62] documented emergent abilities at critical model sizes.

Key Takeaway

Training-time scaling follows precise power laws with phase transitions. Analogous "context scaling laws"—relationships $A(n)$ characterizing context-length regimes—are conjectural but motivated by the non-monotonic curves in S2. No formal context scaling law has been established.

8. Inductive Bias in Neural Models (S5)

Mitchell [35] proved that bias is necessary for generalization. Battaglia et al. [4] taxonomized relational biases across architectures. Soudry et al. [55] showed gradient descent converges to the max-margin solution without explicit regularization. Neyshabur et al. [37, 38] demonstrated that implicit optimization regularization controls generalization.

Xie et al. [64] proved ICL narrows the posterior over latent concepts:

$$p_\theta(y \mid x, C) = \int p_\theta(y \mid x, z)\, p_\theta(z \mid C)\, dz, \tag{13}$$

with $\mathrm{H}(Z \mid C_{1:k})$ decreasing in k (supporting D1 decrease). Garg et al. [12] showed ICL of linear functions matches least-squares; sparse functions match Lasso. Ahuja et al. [2] found Bayesian-predictor approximation. Li et al. [30] provided generalization bounds via algorithm stability. Si et al. [51] revealed systematic biases under underspecified demonstrations.

Key Takeaway

ICL-as-Bayesian-inference provides the strongest theoretical grounding: context concentrates $p_\theta(z \mid C)$, directly decreasing D1 and D4. The framework predicts monotonic improvement with context quality but does not yet account for the overload regime where D_1 increases.

9. Robustness and Prompt Sensitivity (S6)

Sclar et al. [48] showed up to 76-point accuracy differences from meaning-preserving format changes in LLaMA-2-13B ($Q = 2$; D2 $\gg 0$). Zhao et al. [67] identified majority-label, recency, and common-token biases and proposed contextual calibration (+30% absolute, $Q = 2$). Lu et al. [34] showed example order bridges random-guess to SOTA performance. Zhu et al. [68] demonstrated vulnerability to adversarial prompts across 4,788 perturbations. Zhuo et al. [69] found that model confidence correlates with robustness.

Key Takeaway

Semantically equivalent contexts yield dramatically different performance (D2 $\gg 0$, supporting H5), but all evidence is observational ($Q \leq 2$). Controlled manipulation of "bias" independently of content has not been achieved.

10. Medical QA as Benchmark Domain (S7)

We include medical QA solely as a high-stakes benchmark domain that demonstrates context-sensitivity effects; no clinical deployment claims are made.

Primary benchmarks: MedQA [21], MMLU medical subsets [14], PubMedQA [20], MedMCQA [44]. Singhal et al. [52] achieved 67.6% on MedQA; Singhal et al. [53] reached 86.5% with Med-PaLM 2. Nori et al. [40] showed Medprompt achieves >90% without domain-specific training ($Q = 2$)—context engineering substitutes for model-level modification.

Vishwanath et al. [58] found clinical distractors reduce accuracy by 17.9%, with RAG adding 10.3% degradation rather than mitigating it ($Q = 2$; D3 > 0). Li et al. [31] showed that prompting LLMs to ask clarifying questions degrades reasoning quality.

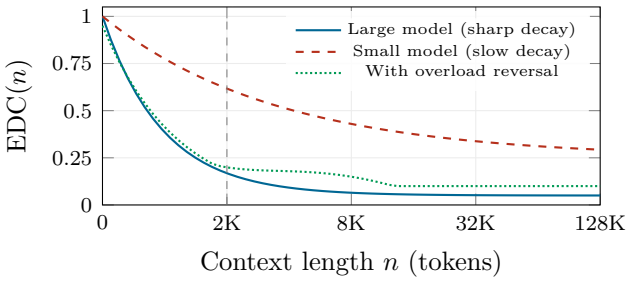

Figure 5: Entropy decay curves (**conceptual schematic**). Larger models exhibit sharper D1 transitions. The overload regime ($n > n_c$) produces non-monotonic D1 profiles. Based on patterns described in the EDC framework [10].

Key Takeaway

Medical QA benchmarks confirm context engineering can match specialist training, but LLMs remain distractor-vulnerable (D3 > 0). RAG can worsen performance, consistent with non-monotonic effects.

11. Entropy Reduction and Search-Space Constriction (S8)

Xie et al. [64] provides the theoretical foundation: ICL concentrates $p_\theta(z \mid C)$, implying D1 decreases with context. Kadavath et al. [22] showed calibrated models produce concentrated distributions on answerable questions. Kuhn et al. [26] introduced semantic entropy, demonstrating that effective context reduces meaning-level uncertainty even when token-level entropy remains high ($Q = 3$). Farquhar et al. [11] used semantic entropy for hallucination detection ($Q = 3$).

The EDC framework [10] defines:

$$\mathrm{EDC}(n) = \frac{\mathrm{H}_\theta(Y \mid x, C_{1:n})}{\mathrm{H}_\theta(Y \mid x)} = \frac{D_1(C_{1:n}, x)}{D_1(\emptyset, x)}, \quad (14)$$

finding that entropy decays with context, larger models show sharper transitions, and profiles vary by task. Jia et al. [19] operationalized EDC for training data selection.

Key Takeaway

Semantic entropy provides the most direct measurement of D1 and D4 as functions of context. EDC reveals model-capacity-dependent decay profiles, supporting H3. Evidence quality is strong ($Q = 3$ for Farquhar et al. and Kuhn et al.).

12. Quantitative Evidence Synthesis

Descriptive effect pooling. Formal meta-analysis is infeasible because studies use heterogeneous metrics, models, and length definitions. For the five highest-quality context-length studies ($Q = 3$: Du et al. 9, Hsieh et al. 18, Kuratov et al.

Table 3: Evidence matrix (12 key studies). Dir.: ↑ improvement, ↓ degradation, ↕ non-monotonic. Manip. type: P = padding, R = retrieval, D = distractor, Pos = position, BF = budget forcing, Fmt = format. Q: study-quality score (Section 2.1).

Study	Task	Model(s)	Manip.	Dir.	Magnitude	Q	Key Control Condition
Du et al. '25	Math/QA/Code	5 LLMs	P, R	↓	13.9–85%	3	Perfect retrieval; whitespace padding
Levy et al. '24	Reasoning	Multiple	P	↓	Significant	3	FLenQA dataset; controlled padding
Liu et al. '24	QA (NQ)	GPT-3.5+	Pos	↕	>30%	3	Position manipulation; fixed length
Cuconasu '24	RAG QA	Multiple	R	↕	+35%	2	Random vs. scored documents
Muennighoff '25	MATH	s1-32B	BF	↕	+27%	3	Budget forcing (append/terminate)
Chen et al. '24	Easy math	o1-like	BF	↓	Varied	2	Overthinking on simple problems
Shi et al. '23	Arithmetic	Multiple	D	↓	>70%	2	Single irrelevant sentence
Vishwanath '25	Med. QA	Med. LLMs	D	↓	17.9%	2	Clinical distractors; RAG adds 10.3%
Nori et al. '23b	MedQA	GPT-4	R	↑	27% err. red.	2	Medprompt context engineering
Sclar et al. '24	Classif.	LLaMA-2	Fmt	↕	76 pts	2	Meaning-preserving format changes
Hsieh et al. '24	RULER	17 LLMs	P	↓	Varied	3	50% fail at claimed context length
Snell et al. '24	Problem solv.	Various	BF	↑	4× eff.	3	Compute-optimal allocation

27, Levy et al. 28, Liu et al. 33), the approximate median degradation per doubling of context length is 8–15% for math/reasoning and 5–10% for factual QA. These are descriptive pooled ranges, not formal effect sizes, and should be interpreted with caution.

Heterogeneity. Effect magnitudes vary substantially: Du et al. report 13.9–85% degradation (a 6× range), reflecting that coding tasks are more context-sensitive than factual QA. Cross-study comparison is limited by differences in model family, task type, context-length range, and manipulation protocol.

Publication bias. The literature likely overrepresents striking findings (large degradation, surprising non-monotonicity) relative to null results. Studies finding modest monotonic improvements are less publishable. This may inflate reported magnitudes, though the *direction* of effects (non-monotonicity exists) is robust given the $Q = 3$ controlled designs of Du et al. [9] and Levy et al. [28].

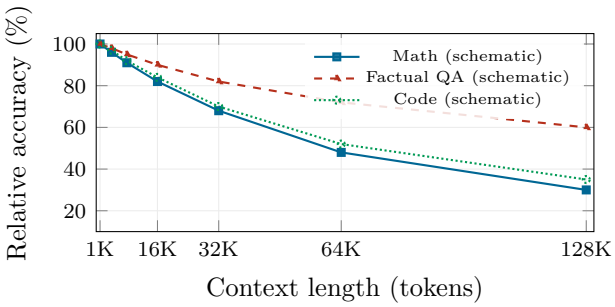

Figure 6: Context-length degradation under perfect retrieval (**conceptual schematic** reflecting reported ranges in Du et al. 2025, not digitized data). Math and coding tasks degrade faster than factual QA.

13. Synthesis: Four Convergent Patterns

Pattern 1: Non-monotonic accuracy is robust and general. Du et al. ($Q = 3$), Levy et al. ($Q = 3$), and Liu et al. ($Q = 3$) collectively demonstrate non-monotonicity across models, tasks, and experimental controls. The whitespace-padding result of Du et al. isolates length as a causal variable (D2 > 0 with C_{irr} = whitespace).

Table 4: Structural parallel between classical and context-induced inductive bias.

Property	**Classical**	**Context-Induced**
Restriction mechanism	Architecture, λ	Length, structure
Non-monotonicity	Bias–variance tradeoff	Accuracy–length curve
Optimization parallel	GD implicit bias	ICL Bayesian inference
IT framework	MDL	IB tradeoff
Measurable signature	Generalization gap	D1 (semantic entropy)
Tunable parameter	λ (regularization)	n (context length)

Table 5: Proposed mechanisms for context-induced effects. Status: **E** = empirical support, **T** = theoretical only, **C** = conjectural. Evidence grade from mean Q of supporting studies.

Mechanism	**St.**	$\bar{Q}$	**Key Citation**
Attention dilution	E	3.0	Liu et al. '24
Posterior narrowing	E	2.0	Xie et al. '22
Retrieval noise	E	2.0	Cuconasu et al. '24
Repr. interference	C	—	Inferred: Du et al. '25
Compression overhead	T	—	IB framework
Positional enc. decay	E	3.0	Liu et al. '24

Pattern 2: Context and reasoning depth interact predictably. Budget forcing yields +27% ($Q = 3$). Medical distractors degrade accuracy by 17.9% ($Q = 2$). The difficulty-dependent optimum $n_r^*(d)$ is supported by both CoT and overthinking evidence.

Pattern 3: Information-theoretic measurements confirm search-space narrowing. Xie et al. (2022) provides theory (posterior concentration); Farquhar et al. ($Q = 3$) and Kuhn et al. ($Q = 3$) provide measurement tools (semantic entropy, D1); the EDC framework provides empirical profiles (D1 as function of n).

Pattern 4: Formal parallels to classical inductive bias hold. Mitchell (1980) establishes bias necessity; the implicit-bias literature shows optimization creates bias without explicit regularization; the ICL literature shows context creates functional equivalents of classical estimators. Table 4 summarizes the parallel.

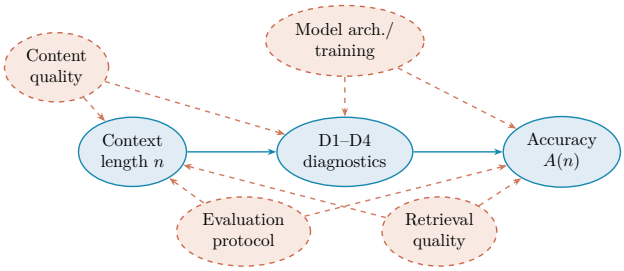

Figure 7: Causal DAG for context-as-bias claims. Solid blue arrows: hypothesized path ($n \to$ diagnostics $\to$ accuracy). Dashed red arrows: confounding paths. Du et al. (2025) partially close the retrieval and content backdoors via perfect retrieval and whitespace controls; model architecture and evaluation protocol confounds remain open. Full identifiability of the $n \to A$ effect is *not* achievable from the current evidence base.

Table 6: Evidence grades per hypothesis. $\bar{Q}$: mean study-quality score from Table 3. Thresholds: Strong ($\bar{Q} \geq 2.5$), Moderate ($1.5 \leq \bar{Q} < 2.5$), Weak ($\bar{Q} < 1.5$).

Hyp.	$\bar{Q}$	Grade	Causal	Diag.
H1: Non-mono. $A(n)$	2.8	Strong	Quasi-exp.	D2, D3
H2: Reas.–length	2.7	Strong	Quasi-exp.	D1
H3: Entropy decay	2.3	Moderate	Correl.	D1, D4
H4: IB modul.	1.2	Weak	Theoretical	MI
H5: Prompt sens.	2.0	Moderate	Observ.	D2

14. Confounding Factors and Causal Limitations

Four classes of confounders threaten causal interpretation: (1) *Model architecture and training data* confound cross-model comparisons. (2) *Evaluation protocol* varies across studies. (3) *Retrieval quality* covaries with length when longer contexts contain more retrieved documents. (4) *Content quality* may degrade with length.

Du et al. (2025, $Q = 3$) close confounds (3) and (4) via perfect retrieval and whitespace padding. Muennighoff et al. (2025, $Q = 3$) achieve quasi-experimental control via budget forcing. Most remaining evidence is observational; confounds (1) and (2) are not addressed by any existing study.

15. What We Can and Cannot Conclude

What the evidence supports. Non-monotonic accuracy under controlled context manipulation (H1) is the best-supported finding, with $Q = 3$ quasi-experimental evidence from multiple independent groups. The reasoning–length interaction (H2) is supported by budget-forcing experiments. Entropy-based diagnostics (H3) show capacity-dependent profiles, though not yet under fully controlled context manipulation.

What remains unestablished. The IB modulation hypothesis (H4) is theoretical: no study has measured $\beta^*(n)$ in a transformer. The bias-signature interpretation of prompt sensitivity (H5) is suggestive but observational—the "bias" component has not been isolated from confounded content effects. The term "inductive bias" is used here by analogy with classical learning theory; we have not proven that context-length effects arise from the same mechanisms as architectural or optimization biases.

Methodological limitations of this review. Single-reviewer screening introduces potential selection bias. The quality rubric, while systematic, involves judgment calls in scoring criteria (b) and (c). The 72-paper corpus, though covering eight streams, is not exhaustive; we may have missed relevant work in adjacent fields (e.g., cognitive science of anchoring, attention mechanism interpretability). The descriptive effect ranges reported in Section 12 are not formal meta-analytic estimates and should not be treated as such.

16. Discussion and Open Problems

The evidence reviewed here establishes that context length exhibits hallmarks of an inductive bias: it constrains the output distribution (D1, D4 decrease), creates non-monotonic accuracy curves (D2 > 0 under irrelevant perturbation), anchors on salient distractors (D3 > 0), and interacts with reasoning depth in a difficulty-dependent manner. Four patterns converge across eight streams, with the two strongest hypotheses (H1, H2) supported by $Q = 3$ quasi-experimental designs.

Key open challenges include establishing the functional form of $A(n)$ (context scaling laws), characterizing the phase transition at n_c, directly measuring the IB tradeoff, and extending findings beyond transformer architectures.

17. Conclusion

This structured review synthesized 72 publications across eight research streams to evaluate the thesis that context length functions as an implicit inductive bias in LLMs. We contributed four operational diagnostics (D1–D4) grounded in the predictive distribution $p_\theta(y \mid x, C)$, five testable hypotheses with explicit falsification conditions, a quality-rubric-graded evidence matrix, causal analysis identifying confounding structure, and a prioritized research agenda.

Two hypotheses receive Strong evidence grades

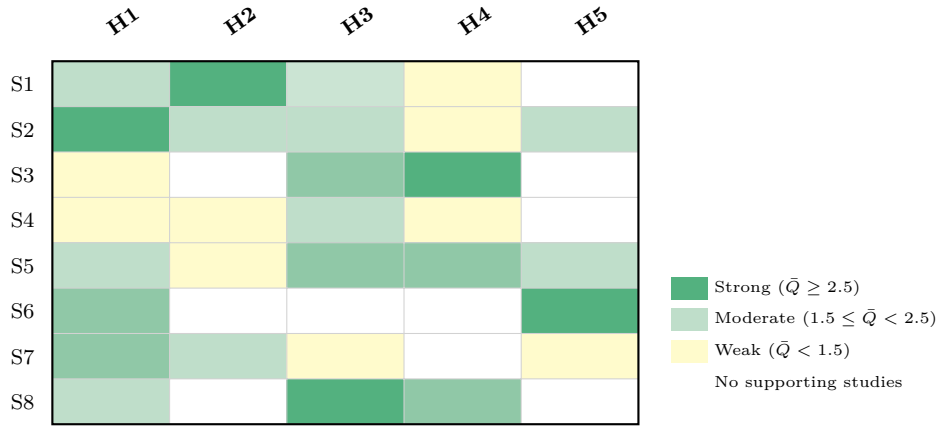

Figure 8: Evidence-map heatmap. Each cell's shading reflects the mean study-quality score ($\bar{Q}$) of papers in that research stream supporting that hypothesis, computed from Table 3. This is a rubric-derived computation, not a subjective assessment.

Table 7: Open problems ranked by expected research value, with proposed experimental protocols and resource estimates.

#	Problem	Description	Proposed Experiment	Effort
1	Context scaling laws	Establish power-law or piecewise $A(n)$ with regime transitions	5+ models × 10 lengths × 5 tasks; perfect retrieval + whitespace controls; measure D1–D4	Med
2	Phase transitions	Identify critical n_c where $\partial A/\partial n$ changes sign	Fine-grained sweeps (256-token steps); EDC inflection measurement; D1 discontinuity detection	Med
3	$\beta^*(n)$ measurement	Direct IB parameter estimation as function of context	Per-layer probing; MINE [5] for $\mathrm{I}(X;T)$ and $\mathrm{I}(T;Y)$ under varying n	High
4	Optimal allocation	Adaptive strategies for context budgeting	Difficulty-aware allocation; compare against fixed-length baselines across task difficulties	Med
5	Distractor mechanism	Disambiguate attention dilution, repr. interference, and posterior corruption	Ablation: attention maps, intermediate representations, output distributions ± controlled distractors	High
6	Cross-architecture	Test context-as-bias on Mamba/SSM and hybrid architectures	Replicate Du et al. (2025) whitespace-padding protocol on non-transformer models	High

($\bar{Q} \geq 2.5$): non-monotonic accuracy (H1) and reasoning–length interaction (H2). Two receive Moderate grades: entropy decay regimes (H3) and prompt sensitivity (H5). One remains theoretical: IB modulation (H4). The convergence of test-time compute scaling, robustness analysis, and information-theoretic measurement establishes context length as a first-class axis of model behavior warranting dedicated investigation alongside architecture, training data, and optimization.

References

[1] A. Achille and S. Soatto. Emergence of invariance and disentanglement in deep representations. *JMLR*, 19(50):1–34, 2018.

[2] K. Ahuja, M. Panwar, and N. Goyal. In-context learning through the Bayesian prism. In *ICLR*, 2024.

[3] Y. Bai et al. LongBench: A bilingual, multi-

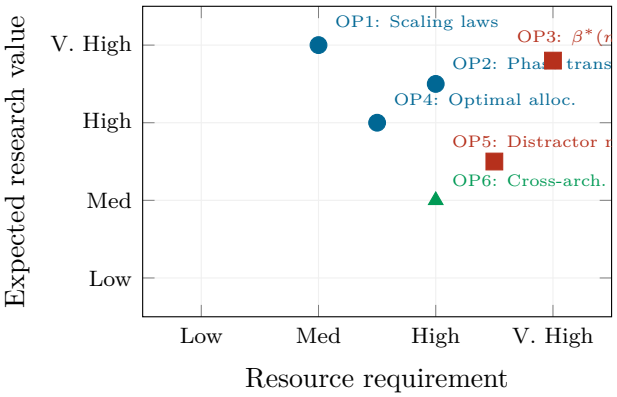

Figure 9: Open-problems roadmap: expected value vs. resource cost. Marker shapes: circles (single-lab feasible), squares (multi-lab), triangles (community effort). OP1 offers the highest value-to-cost ratio.

task benchmark for long context understanding. In *ACL*, 2024.

[4] P. W. Battaglia et al. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261*, 2018.

[5] M. I. Belghazi et al. MINE: Mutual information neural estimation. In *ICML*, PMLR 80:531–540, 2018.

[6] X. Chen et al. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv:2412.21187*, 2024.

[7] A. Clark et al. Unified scaling laws for routed language models. In *ICML*, pp. 4057–4086, 2022.

[8] F. Cuconasu et al. The power of noise: Redefining retrieval for RAG systems. In *SIGIR*, pp. 719–729, 2024.

[9] Y. Du et al. Context length alone hurts LLM performance despite perfect retrieval. In *EMNLP Findings*, 2025.

[10] EDC Framework. Measuring and analyzing intelligence via contextual uncertainty in LLMs. *arXiv:2507.21129*, 2025.

[11] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

[12] S. Garg, D. Tsipras, P. Liang, and G. Valiant. What can transformers learn in-context? In *NeurIPS*, 2022.

[13] D. Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*, 2025.

[14] D. Hendrycks et al. Measuring massive multitask language understanding. In *ICLR*, 2021.

[15] T. Henighan et al. Scaling laws for autoregressive generative modeling. *arXiv:2010.14701*, 2020.

[16] R. D. Hjelm et al. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[17] J. Hoffmann et al. Training compute-optimal large language models. In *NeurIPS*, 2022.

[18] C.-P. Hsieh et al. RULER: What's the real context size of your long-context language models? In *COLM*, 2024.

[19] J. Jia et al. EntropyLong: Effective long-context training via predictive uncertainty. *arXiv:2510.02330*, 2025.

[20] Q. Jin et al. PubMedQA: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, pp. 2567–2577, 2019.

[21] D. Jin et al. What disease does this patient have? *Applied Sciences*, 11(14):6421, 2021.

[22] S. Kadavath et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

[23] G. Kamradt. Needle in a haystack—pressure testing LLMs. GitHub repository, 2023.

[24] J. Kaplan et al. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[25] T. Kojima et al. Large language models are zero-shot reasoners. In *NeurIPS*, pp. 22199–22213, 2022.

[26] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in NLG. In *ICLR*, 2023.

[27] Y. Kuratov et al. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *NeurIPS*, pp. 106519–106554, 2024.

[28] M. Levy, A. Jacoby, and Y. Goldberg. Same task, more tokens: The impact of input length on reasoning performance of LLMs. In *ACL*, pp. 15339–15353, 2024.

[29] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, pp. 9459–9474, 2020.

[30] Y. Li et al. Transformers as algorithms: Generalization and stability in in-context learning. In *ICML*, PMLR 202:19565–19594, 2023.

[31] S. S. Li et al. MEDIQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *NeurIPS*, 2024.

[32] H. Lightman et al. Let's verify step by step. In *ICLR*, 2024.

[33] N. F. Liu et al. Lost in the middle: How language models use long contexts. *TACL*, 12:157–173, 2024.

[34] Y. Lu et al. Fantastically ordered prompts and where to find them. In *ACL*, pp. 8086–8098, 2022.

[35] T. M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

[36] N. Muennighoff et al. s1: Simple test-time scaling. In *EMNLP*, pp. 20275–20321, 2025.

[37] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias. In *ICLR Workshop*, 2015.

[38] B. Neyshabur et al. Geometry of optimization and implicit regularization in deep learning. *arXiv:1705.03071*, 2017.

[39] H. Nori et al. Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375*, 2023a.

[40] H. Nori et al. Can generalist foundation models outcompete special-purpose tuning? *arXiv:2311.16452*, 2023b.

[41] M. Nye et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv:2112.00114*, 2021.

[42] OpenAI. OpenAI o1 system card. *arXiv:2412.16720*, 2024.

[43] OpenAI. OpenAI o3 and o4-mini system card. April 2025.

[44] A. Pal et al. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain QA. In *CHIL*, PMLR 174:248–260, 2022.

[45] P. Sahoo et al. A systematic survey of prompt engineering in LLMs. *arXiv:2402.07927*, 2024.

[46] N. Sardana et al. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *ICML*, pp. 43445–43460, 2024.

[47] A. M. Saxe et al. On the information bottleneck theory of deep learning. *J. Stat. Mech.*, 2019(12):124020, 2019.

[48] M. Sclar et al. Quantifying language models' sensitivity to spurious features in prompt design. In *ICLR*, 2024.

[49] F. Shi et al. Large language models can be easily distracted by irrelevant context. In *ICML*, PMLR 202:31210–31227, 2023.

[50] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

[51] C. Si et al. Measuring inductive biases of in-context learning with underspecified demonstrations. *arXiv:2305.13299*, 2023.

[52] K. Singhal et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023a.

[53] K. Singhal et al. Towards expert-level medical question answering with LLMs. *Nature Medicine*, 30:1114–1122, 2024.

[54] C. Snell et al. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*; In *ICLR*, 2025.

[55] D. Soudry et al. The implicit bias of gradient descent on separable data. *JMLR*, 19(1):2822–2878, 2018.

[56] Y. Sui et al. Stop overthinking: A survey on efficient reasoning for LLMs. *arXiv:2503.16419*, 2025.

[57] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Allerton Conference*, pp. 368–377, 1999.

[58] K. Vishwanath et al. Medical large language models are easily distracted. *arXiv:2504.01201*, 2025.

[59] X. Wang et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[60] Y. Wang, P. Li, and Y. Yang. Visual transformer with differentiable channel selection: An IB inspired approach. In *ICML*, PMLR 235:50840–50858, 2024.

[61] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pp. 24824–24837, 2022a.

[62] J. Wei et al. Emergent abilities of large language models. *TMLR*, 2022b.

[63] Y. Wu et al. Inference scaling laws: An empirical analysis of compute-optimal inference. *arXiv:2408.00724*; In *ICLR*, 2025.

[64] S. M. Xie et al. An explanation of in-context learning as implicit Bayesian inference. In *ICLR*, 2022.

[65] S. Yao et al. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023.

[66] E. Zelikman et al. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, pp. 15476–15488, 2022.

[67] T. Z. Zhao et al. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

[68] K. Zhu et al. PromptRobust: Towards evaluating the robustness of LLMs on adversarial prompts. *arXiv:2306.04528*, 2023.

[69] J. Zhuo et al. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of EMNLP*, pp. 1950–1976, 2024.

A. Executive Summary for Non-Specialists

Large language models process text through a "context window." Conventional wisdom holds that more context improves performance. This review of 72 studies demonstrates otherwise: the *amount* of context—independent of its quality—fundamentally changes how models reason. Performance follows a non-monotonic pattern: too little context leaves the model guessing; optimal context focuses reasoning; too much degrades performance, even when all information is correct and relevant, or when extra tokens consist entirely of whitespace.

We argue context length functions as an "inductive bias"—a structural constraint shaping what solutions the model considers. We define four measurable diagnostics for detecting this bias effect and state five testable hypotheses with explicit conditions under which they would be falsified. Practical implication: rather than maximizing context, practitioners should match context length to task difficulty.

B. Data Extraction Schema and Rubric Examples

Extraction fields (CSV columns). `bib_key`, `year`, `streams` (S1–S8, comma-separated), `task_domain`, `model_family`, `context_min_tokens`, `context_max_tokens`, `manipulation_type` (P/R/D/Pos/BF/Fmt), `effect_direction` (↑/↓/↕), `effect_magnitude`, `control_conditions`, `Q_controlled` (0/1), `Q_clarity` (0/1), `Q_data` (0/1), `Q_breadth` (0/1), `Q_uncertainty` (0/1), `Q_total` (0–3), `hypotheses_supported` (H1–H5).

Rubric scoring examples. *Du et al. (2025),* $Q = 3$: (a) Controlled manipulation: ✓ (whitespace padding, perfect retrieval). (b) Protocol clarity: ✓ (model versions, prompts, decoding parameters fully specified). (c) Data/code: ✓ (FLenQA released). (d) Breadth: ✓ (5 model families, 3 task types). (e) Uncertainty: ✓ (multiple runs). Score: $5/5 \rightarrow Q = 3$.

Sclar et al. (2024), $Q = 2$: (a) Controlled: ✓ (meaning-preserving format changes). (b) Clarity: ✓ (prompts specified). (c) Data: partial (some code). (d) Breadth: mixed (multiple models, single task type). (e) Uncertainty: × (no CIs). Score: $2.5/5 \rightarrow Q = 2$.

Kadavath et al. (2022), $Q = 1$: (a) Controlled: × (observational calibration study). (b) Clarity: ✓. (c) Data: ×. (d) Breadth: ✓ (multiple models). (e) Uncertainty: ×. Score: $2/5 \rightarrow Q = 1$.

Included papers. All 72 included papers are listed in the bibliography. Numeric values cited in Table 3 are drawn directly from original papers (tables, abstracts, or main text). No values were digitized from plots; all figures in this review are labeled as conceptual schematics.

C. Anticipated Critiques

C1: "The thesis is unfalsifiable." Each hypothesis (Section 3.3) specifies an explicit falsification condition. H1 would be falsified by monotonic improvement under the whitespace-padding protocol of Du et al. [9].

C2: "No new experiments are presented." Correct. We position this as a structured review with formal synthesis. Contributions include operational diagnostics (D1–D4), testable hypotheses, a quality-graded evidence matrix, causal analysis, and a prioritized experimental agenda.

C3: "H4 is speculative." We agree and rate it Weak ($\bar{Q} = 1.2$). H4 is presented as a motivated conjecture with a concrete falsification condition, not a demonstrated result.

C4: "Publication bias inflates effect sizes." Acknowledged in Section 12. The direction of effects is robust given $Q = 3$ controlled designs; magnitudes should be interpreted cautiously.

C5: "The diagnostic framework is not validated." D1 and D4 are directly measurable from model outputs. D2 requires paired context conditions. D3 requires controlled distractors. We propose

validation experiments in Table 7 (OP1).

Repository structure.

```
context-bias-review/
  data/evidence_matrix.csv
  data/paper_metadata.csv
  data/quality_scores.csv
  figures/generate_archetypes.py
  figures/generate_edc.py
  figures/generate_heatmaps.py
  analysis/compute_pooled_ranges.py
  analysis/compute_evidence_grades.py
  paper/main.tex
  paper/references.bib
  README.md
```