Phase 2 Corpus Sources: LLM Faithfulness & Citation Correctness

Research Question: How do different prompting strategies affect the faithfulness and citation correctness of LLM-generated research answers?

Total Sources: 28 (16 peer-reviewed papers, 4 benchmarks, 3 surveys, 3 technical reports/frameworks, 2 empirical/critique studies)

Bucket 1: Faithfulness & Groundedness in LLMs (6 sources)

S01 — Survey on Hallucination in LLMs (Huang et al.)

Field	Value
source_id	S01
title	A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions
authors	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu
year	2023 (updated 2024)
type	Survey (peer-reviewed)
venue	ACM Transactions on Information Systems (TOIS)
link/DOI	https://arxiv.org/abs/2311.05232 / DOI: 10.1145/3703155
relevance	Foundational survey defining factuality vs. faithfulness hallucination taxonomy. Directly supports your definitions of faithfulness and provides the conceptual framework for evaluating when LLM outputs diverge from provided context.

S02 — FActScore (Min et al.)

Field	Value
source_id	S02
title	FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Field	Value
authors	Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, Hannaneh Hajishirzi
year	2023
type	Benchmark paper (peer-reviewed)
venue	EMNLP 2023
link/DOI	https://arxiv.org/abs/2305.14251 / DOI: 10.18653/v1/2023.emnlp-main.741
relevance	Introduces atomic fact decomposition as a unit for measuring factual precision — directly relevant to your pipeline's groundedness evaluation metric. The automated estimator approach can inform your own evaluation methodology.

S03 — TruthfulQA (Lin et al.)

Field	Value
source_id	S03
title	TruthfulQA: Measuring How Models Mimic Human Falsehoods
authors	Stephanie Lin, Jacob Hilton, Owain Evans
year	2022
type	Benchmark paper (peer-reviewed)
venue	ACL 2022
link/DOI	https://arxiv.org/abs/2109.07958 / DOI: 10.18653/v1/2022.acl-long.229
relevance	Key benchmark for measuring LLM truthfulness. Its finding that larger models are often less truthful (inverse scaling) directly informs Sub-Question A about hallucination without grounding. Provides methodology for adversarial question design.

S04 — FAVA: Fine-grained Hallucination Detection (Mishra et al.)

Field	Value
source_id	S04

Field	Value
title	Fine-grained Hallucination Detection and Editing for Language Models
authors	Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, Hannaneh Hajishirzi
year	2024
type	Empirical study (peer-reviewed)
venue	COLM 2024
link/DOI	https://arxiv.org/abs/2401.06855
relevance	Introduces a 6-type hallucination taxonomy and FavaBench benchmark. Directly relevant to Sub-Question E (types of citation/faithfulness errors). Shows ChatGPT and Llama2 hallucinate in 60-75% of information-seeking outputs.

S05 — Measuring and Improving Faithfulness of Chain-of-Thought (Paul et al.)

Field	Value
source_id	S05
title	Measuring and Improving Faithfulness of Chain-of-Thought Reasoning
authors	Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, Boi Faltings
year	2024
type	Empirical study (peer-reviewed)
venue	EMNLP 2024 Findings
link/DOI	https://aclanthology.org/2024.findings-emnlp.882.pdf
relevance	Empirically measures whether CoT reasoning steps actually influence model outputs via causal mediation analysis. Directly relevant to Sub-Question A (whether reasoning chains are faithful to evidence) and Sub-Question C (prompting strategies for faithfulness).

S06 — Assessing Faithfulness of LLM-generated Feedback (Jia et al.)

Field	Value
source_id	S06
title	On Assessing the Faithfulness of LLM-generated Feedback on Student Assignments
authors	Qinjin Jia, Jialin Cui, Ruijie Xi, Chengyuan Liu, Parvez Rashid, Ruochi Li, Edward Gehringer
year	2024
type	Empirical study (peer-reviewed)
venue	EDM 2024 (17th International Conference on Educational Data Mining)
link/DOI	https://files.eric.ed.gov/fulltext/ED675643.pdf
relevance	Applied faithfulness evaluation comparing data-driven (BART fine-tuned, 27.1% hallucination) vs. prompt-driven (ChatGPT-4 few-shot, 23.5% hallucination) systems. Finds intrinsic hallucinations dominate in fine-tuned models while extrinsic hallucinations dominate in prompt-driven systems. Tests NLI-based and ChatGPT-based hallucination measurement with best F1 ≈ 72%. Bridges educational/research assistance domain — closest to your pipeline's use case. Relevant to Sub-Questions A, B, and C.

Bucket 2: Citation Correctness & Attribution (5 sources)

S07 — ALCE Benchmark (Gao et al.)

Field	Value
source_id	S07
title	Enabling Large Language Models to Generate Text with Citations
authors	Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen
year	2023
type	Benchmark paper (peer-reviewed)
venue	EMNLP 2023
link/DOI	https://arxiv.org/abs/2305.14627 / DOI: 10.18653/v1/2023.emnlp-main.398

Field	Value
relevance	Core benchmark for your research — the first reproducible benchmark for evaluating LLM citation quality across fluency, correctness, and citation dimensions. Directly defines citation recall and precision metrics you can adopt. Shows best models lack complete citation support 50% of the time.

S08 — Measuring Attribution in NLG / AIS Framework (Rashkin et al.)

Field	Value
source_id	S08
title	Measuring Attribution in Natural Language Generation Models
authors	Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, David Reitter
year	2023
type	Framework paper (peer-reviewed)
venue	Computational Linguistics, Vol. 49, No. 4 (MIT Press)
link/DOI	https://arxiv.org/abs/2112.12870 / DOI: 10.1162/coli_a_00486
relevance	Defines the Attributable to Identified Sources (AIS) framework — the foundational formalism for your definition of citation correctness. Provides annotation guidelines and a two-stage evaluation pipeline applicable to your own evaluation set design.

S09 — Evaluating Verifiability in Generative Search Engines (Liu et al.)

Field	Value
source_id	S09
title	Evaluating Verifiability in Generative Search Engines
authors	Nelson F. Liu, Tianyi Zhang, Percy Liang
year	2023
type	Empirical study (peer-reviewed)
venue	EMNLP 2023 Findings

Field	Value
link/DOI	https://arxiv.org/abs/2304.09848 / DOI: 10.18653/v1/2023.findings-emnlp.467
relevance	Audits citation recall and precision in real commercial systems (Bing Chat, Perplexity, etc.). Finds only 51.5% of statements are fully supported by citations. Directly informs Sub-Question E (types of citation errors) and provides citation precision/recall methodology you can replicate.

S10 — Chain-of-Thought Improves Text Generation with Citations (AAAI 2024)

Field	Value
source_id	S10
title	Chain-of-Thought Improves Text Generation with Citations in Large Language Models
authors	(Authors from AAAI 2024 proceedings)
year	2024
type	Empirical study (peer-reviewed)
venue	AAAI 2024
link/DOI	https://ojs.aaai.org/index.php/AAAI/article/view/29794
relevance	Directly tests CoT prompting for citation generation on the ALCE benchmark across 6 LLMs. Shows CoT consistently improves citation precision and recall — core evidence for Sub-Questions A and C about how prompting strategies affect citation correctness.

S11 — Survey of LLM Attribution (Li et al.)

Field	Value
source_id	S11
title	A Survey of Large Language Models Attribution
authors	(Authors from HITsz-TMG)
year	2023
type	Survey
venue	arXiv preprint (arXiv:2311.03731)

Field	Value
link/DOI	https://arxiv.org/abs/2311.03731
relevance	Comprehensive survey covering pre-generation, in-generation, and post-generation attribution approaches. Provides a taxonomy of attribution systems and their features, useful for understanding the landscape of citation methods your pipeline should be aware of.

Bucket 3: Prompting Strategies (5 sources)

S12 — Chain-of-Thought Prompting (Wei et al.)

Field	Value
source_id	S12
title	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
authors	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Fei Xia, Quoc Le, Denny Zhou
year	2022
type	Empirical study (peer-reviewed)
venue	NeurIPS 2022
link/DOI	https://arxiv.org/abs/2201.11903
relevance	The foundational paper on chain-of-thought prompting. Directly relevant as one of the prompting strategies you'll evaluate — demonstrates that intermediate reasoning steps improve performance on complex tasks, which may or may not improve faithfulness.

S13 — Hallucination Attribution: Prompting vs. Model Behavior

Field	Value
source_id	S13
title	Survey and Analysis of Hallucinations in Large Language Models: Attribution to Prompting Strategies or Model Behavior
authors	(PMC/NIH published, multi-author)

Field	Value
year	2025
type	Empirical study (peer-reviewed)
venue	PMC (peer-reviewed journal)
link/DOI	https://pmc.ncbi.nlm.nih.gov/articles/PMC12518350/
relevance	Directly addresses your main research question by empirically separating prompt-induced hallucinations from model-intrinsic ones. Tests multiple LLMs with standardized benchmarks and proposes a diagnostic framework — highly relevant to Sub-Questions A and D.

S14 — Comprehensive Survey on Trustworthiness in Reasoning

Field	Value
source_id	S14
title	A Comprehensive Survey on Trustworthiness in Reasoning
authors	(Multi-author, OpenReview)
year	2024
type	Survey (peer-reviewed)
venue	OpenReview (submitted for peer review)
link/DOI	https://openreview.net/pdf?id=Ysslwdjb6L
relevance	Covers reasoning faithfulness specifically — distinguishes between answers being correct vs. the reasoning process being faithful. Clarifies confusion between different definitions of faithfulness in the literature, directly supporting your Definitions section.

S15 — The Decreasing Value of Chain of Thought in Prompting

Field	Value
source_id	S15
title	Prompting Science Report 2: The Decreasing Value of Chain of Thought in Prompting
authors	Lennart Meincke, Ethan R. Mollick, Lilach Mollick, Dan Shapiro

Field	Value
year	2025
type	Technical report
venue	The Wharton School Research Paper (SSRN)
link/DOI	https://ssrn.com/abstract=5285532
relevance	Critique paper showing CoT effectiveness varies by model type and task — reasoning models gain minimal benefit from explicit CoT. Provides important nuance for Sub-Question C about whether prompting strategies work consistently across models.

S16 — Towards Faithful Model Explanation in NLP (Lyu et al.)

Field	Value
source_id	S16
title	Towards Faithful Model Explanation in NLP: A Survey
authors	Qing Lyu et al.
year	2024
type	Survey (peer-reviewed)
venue	Computational Linguistics (ACL)
link/DOI	https://aclanthology.org/2024.cl-2.6.pdf
relevance	Reviews 110+ explanation methods through the lens of faithfulness. Provides rigorous definitions and evaluation principles for faithfulness that complement your operational definitions. Relevant to understanding whether LLM explanations accurately reflect reasoning.

Bucket 4: Hallucination Taxonomy & Detection (3 sources)

S17 — LLM Hallucination: A Comprehensive Survey (Alansari & Luqman)

Field	Value
source_id	S17

Field	Value
title	Large Language Models Hallucination: A Comprehensive Survey
authors	Aisha Alansari, Hamzah Luqman
year	2025
type	Survey
venue	arXiv preprint (arXiv:2510.06265)
link/DOI	https://arxiv.org/abs/2510.06265
relevance	Most recent comprehensive hallucination survey covering the full LLM development lifecycle. Provides updated taxonomy of detection and mitigation approaches, and reviews current benchmarks — useful for ensuring your evaluation methodology is current.

S18 — Walk the Talk? Measuring Faithfulness of LLM Explanations

Field	Value
source_id	S18
title	Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations
authors	(OpenReview submission)
year	2024
type	Empirical study
venue	OpenReview
link/DOI	https://openreview.net/forum?id=4ub9gpx9xw
relevance	Introduces a novel method for measuring explanation faithfulness by testing whether the concepts LLMs claim are influential actually are. Provides a rigorous definition of faithfulness relevant to Sub-Question A.

S19 — Comprehensive Survey of Faithfulness Evaluation Methods

Field	Value
source_id	S19

Field	Value
title	A Comprehensive Survey of Faithfulness Evaluation Methods
authors	(RANLP 2025 proceedings)
year	2025
type	Survey (peer-reviewed)
venue	RANLP 2025
link/DOI	https://acl-bg.org/proceedings/2025/RANLP%202025/pdf/2025.ranlp-1.74.pdf
relevance	Surveys faithfulness evaluation methods specifically, including fact-based, classifier-based, QA-based, and LLM-based approaches. Directly informs your choice of evaluation metrics for the RAG pipeline.

S20 — The Dawn After the Dark: Empirical Study on Factuality Hallucination (Li et al.)

Field	Value
source_id	S20
title	The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models
authors	Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen
year	2024
type	Empirical study (peer-reviewed)
venue	ACL 2024 (Volume 1: Long Papers)
link/DOI	https://aclanthology.org/2024.acl-long.586/
relevance	Comprehensive empirical study spanning hallucination detection, source, and mitigation across pre-training, SFT, RLHF, and inference stages. Introduces HaluEval 2.0 benchmark (8,770 questions, 5 domains) and a 6-type factuality hallucination taxonomy (entity-error, relation-error, incompleteness, outdatedness, overclaim, unverifiability). Key findings: retrieval augmentation significantly reduces hallucinations; CoT helps larger models but hurts smaller ones; prompt design (task descriptions, in-context demos) affects hallucination rates. Directly relevant to Sub-Questions A, B, and C.

S21 — HaluEval: Hallucination Evaluation Benchmark (Li et al.)

Field	Value
source_id	S21
title	HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models
authors	Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen
year	2023
type	Benchmark paper (peer-reviewed)
venue	EMNLP 2023
link/DOI	https://aclanthology.org/2023.emnlp-main.397.pdf
relevance	Introduces a 35,000-sample hallucination evaluation benchmark across QA, dialogue, and summarization with a sampling-then-filtering generation pipeline. Finds ChatGPT hallucinates in ~19.5% of responses, and that knowledge retrieval boosts recognition accuracy (62.59% → 76.83% in QA) while CoT has mixed effects. Tests hallucination patterns (comprehension, factualness, specificity, inference) showing topic-sensitive hallucination. Relevant to Sub-Questions A, B, and E.

Bucket 5: RAG Pipeline Design (4 sources)

S22 — RAG: Original Paper (Lewis et al.)

Field	Value
source_id	S22
title	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
authors	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela
year	2020
type	Foundational paper (peer-reviewed)
venue	NeurIPS 2020

Field	Value
link/DOI	https://arxiv.org/abs/2005.11401
relevance	The original RAG paper — foundational to your entire pipeline architecture. Defines the retrieve-then-generate paradigm, parametric vs. non-parametric memory, and demonstrates RAG produces more factual language than parametric-only baselines.

S23 — Self-RAG (Asai et al.)

Field	Value
source_id	S23
title	Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection
authors	Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, Hannaneh Hajishirzi
year	2023
type	Empirical study (peer-reviewed)
venue	ICLR 2024
link/DOI	https://arxiv.org/abs/2310.11511
relevance	Introduces reflection tokens for adaptive retrieval and self-critique — demonstrates significant gains in factuality and citation accuracy over standard RAG. Relevant to your enhancement options (confidence scoring, trust behavior) and Sub-Question C.

S24 — RAG Survey (Gao et al.)

Field	Value
source_id	S24
title	Retrieval-Augmented Generation for Large Language Models: A Survey
authors	Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang
year	2024
type	Survey

Field	Value
venue	arXiv preprint (arXiv:2312.10997)
link/DOI	https://arxiv.org/abs/2312.10997
relevance	Comprehensive RAG survey covering Naive, Advanced, and Modular RAG paradigms. Provides the technical context for your pipeline design decisions (chunking, retrieval, augmentation) and introduces evaluation frameworks and benchmarks.

S25 — R2-MGA: Verifiable Text Generation with Generative Agents

Field	Value
source_id	S25
title	Towards Verifiable Text Generation with Generative Agent
authors	(AAAI 2025 proceedings)
year	2025
type	Empirical study (peer-reviewed)
venue	AAAI 2025
link/DOI	https://ojs.aaai.org/index.php/AAAI/article/view/34599
relevance	Proposes an agent-based approach to citation generation that achieves +154.7% citation quality improvement on ALCE. Demonstrates that retrieval of best-matched demonstrations significantly improves citation recall and precision — relevant to your enhancement choices.

Bucket 6: Supporting Sources (3 sources)

S26 — DeepEval Faithfulness Metric Documentation

Field	Value
source_id	S26
title	Faithfulness Metric — DeepEval Documentation
authors	Confident AI (DeepEval team)

Field	Value
year	2024
type	Technical documentation
venue	deepeval.com
link/DOI	https://deepeval.com/docs/metrics-faithfulness
relevance	Practical implementation guide for faithfulness evaluation using LLM-as-a-judge. Provides the QAG scorer methodology and code examples you can adapt for your pipeline's evaluation component.

S27 — RAGAS Faithfulness Metric Documentation

Field	Value
source_id	S27
title	Faithfulness — RAGAS Documentation
authors	RAGAS team
year	2024
type	Technical documentation
venue	docs.ragas.io
link/DOI	https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/faithfulness/
relevance	Practical faithfulness scoring implementation using claim extraction and NLI verification. Includes integration with Vectara's HHEM-2.1-Open classifier — provides a concrete evaluation approach for your RAG pipeline.

S28 — Confident AI LLM Evaluation Metrics Guide

Field	Value
source_id	S28
title	LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide
authors	Confident AI

Field	Value
year	2024
type	Technical guide
venue	confident-ai.com
link/DOI	https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation
relevance	Comprehensive guide to RAG evaluation metrics covering faithfulness, answer relevance, context precision/recall, and hallucination metrics. Useful for selecting your additional evaluation metric beyond groundedness.

Summary by Type

Type	Count	Source IDs
Peer-reviewed papers	16	S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S12, S20, S21, S22, S23, S25
Benchmarks (within above)	5	S02, S03, S07, S08, S21
Surveys	5	S01, S11, S14, S17, S24
Empirical/Critique studies	4	S13, S15, S18, S19
Technical docs/guides	3	S26, S27, S28

Summary by Relevance to Sub-Questions

Sub-Question	Most Relevant Sources
A. Hallucination without grounding	S01, S03, S04, S05, S06, S13, S18, S20, S21
B. Task-dependent error rates	S04, S06, S07, S09, S10, S20, S21
C. Uncertainty acknowledgment under prompting	S05, S06, S12, S13, S15, S20, S23
D. Cross-model consistency	S03, S04, S07, S13
E. Types of citation errors	S04, S07, S08, S09, S19, S20, S21

Next Steps

1. **Download PDFs** for S01–S25 from the arXiv/ACL links above into `data/raw/`
2. **Save HTML snapshots** for S26–S28 (technical docs) into `data/raw/`
3. **Convert this document** into your `data/manifest.csv` or `manifest.json`
4. Write a short note in your README: "Sources were selected manually via systematic search across Google Scholar and arXiv, targeting peer-reviewed papers on LLM faithfulness, citation generation, hallucination taxonomy, prompting strategies, and RAG design published 2020–2025."