INDUCTIVE TRIPLET FINE TUNING FOR SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most reasoning evaluations conflate deduction with induction. We target *inductive* ability; i.e., ampliative inference from noisy evidence; and introduce a (Context, Question, Answer) triplet corpus aligned to ten canonical inductive reasoning (IR) forms (enumeration, statistical generalization/syllogism, analogy, default rules, abduction, Bayesian/Carnapian updates, Mill-style causal inference). The dataset (IR-Triplets) is fairly balanced across forms. We fine-tune ten small language models (SLMs) (0.5B–9B) with parameter-efficient Supervised Fine-Tuning (SFT) and evaluate in a 2×2 design: in-distribution (ID) held-out data and out-of-distribution (OOD) transfer to a different dataset DEER. IR-Triplets dataset yields consistent ID gains in Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (ROUGE-L) (mean absolute ≈ 0.07, ~ 60–70% relative), with large improvements for several models; OOD transfer is heterogeneous but frequently positive (e.g., Gemma2-2B, Llama-8B). Post-hoc spectral diagnostics show strong compression: spectral tail index and stable rank typically drop by ~ 45–80% and ~ 44–55%, respectively. Ordinary Least Squares (OLS) analyses clarify that model size strongly predicts spectral compression, while ROUGE-L gains are not a significant predictor once size is controlled; conversely, spectral deltas do not significantly explain ROUGE-L gains in the reverse regression with this sample size. Overall, IR-Triplets dataset reliably improves text-level fidelity and reorganizes capacity toward lower-rank, heavier-tailed representations, but the magnitude of ROUGE-L improvement does not linearly track the *amount* of global compression, pointing to subspace-level mechanisms as a key direction for OOD robustness. [1]

1 INTRODUCTION

Large language models (LLMs) Xiao & Zhu (2025) have shown strong performance in natural language interaction, renewing interest in their *reasoning* abilities Lee et al. (2024). Yet most evaluations blur two distinct modes: *deductive* reasoning (deriving conclusions from general rules) and *inductive* reasoning (inferring rules, patterns, or decisions from examples under uncertainty) Cheng et al. (2024) . This paper explicitly targets the latter.

This work. We curate a corpus of (Context, Question, Answer) (or (C,Q,A)) , *IR-Triplets*, aligned to ten canonical inductive forms (enumeration, statistical generalization/syllogism, analogy, default rules, abduction, Bayesian/Carnapian updates, and Mill-style causal inference) detailed in §2.1. The dataset is deliberately balanced across forms (N=1,807), enabling analysis that does not hinge on a single pattern of inference. Extraction is schema-constrained with evidence spans and automated consistency checks to curb hallucination and improve auditability. We then fine-tune *small language models* (SLMs) Wang et al. (2024a), motivated by their practicality for agentic AI systems Belcak et al. (2025), and evaluate both in-distribution (ID) and out-of-distribution (OOD) on a dataset produced entirely outside our data-generation process.

Empirical findings. Across ten SLMs (0.5B–9B) trained with parameter-efficient Supervised Fine-Tuning (SFT), we observe consistent ID gains on ROUGE-L Lin (2004) and frequent OOD

[1] Editing disclosure: We used OpenAI's ChatGPT solely for grammar and phrasing. It was not used for ideas, methods, or results; the author takes full responsibility for the content.

improvements. Post-hoc spectral analyses Martin & Mahoney (2021) show strong *spectral compression* (drops in spectral tail index and stable rank). Controlled regressions indicate that *model size* is a primary driver of compression, while ROUGE-L gains do not linearly track global spectral shifts once size is controlled—suggesting that *where* compression occurs (task-relevant subspaces) matters more than its global magnitude.

Related work. ARC Chollet (2019) was introduced to probe human-like fluid intelligence with developer-aware generalization and core-knowledge priors; recent work leverages ARC to structure hypothesis articulation in language and subsequent program synthesis Wang et al. (2024b). To disentangle induction from deduction, Cheng et al. (2024) propose SOLVERLEARNER, separating learning input–output functions from applying them and using counterfactual tasks to probe *pure* induction; reporting near-perfect inductive performance for recent LLMs (e.g., GPT-4) but weaker OOD deductive robustness. Moskvichev et al. (2023) introduce CONCEPTARC, clustering tasks into concept families to test abstraction and generalization; humans far outperform GPT-4 and top ARC-Kaggle systems, revealing persistent gaps. Li et al. (2025) investigates engines that extract inductive rules directly. In parallel, Jin et al. (2025) show that chain-of-thought prompting can *hurt* induction on complex *special* rules (e.g., chess, poker, dice, blackjack) and propose structured fixes (guided decomposition, non-numeric exemplars, strict summarization limits) that recover or improve accuracy. Complementarily, Yang et al. (2024) teach models to infer *natural-language* rules from *natural-language* facts, introducing DEER/DEERLET and the philosophy-inspired COLM framework that filters rules by consistency, reality, generality, and non-triviality. Our work differs in (i) treating induction as supervised mapping from evidence to concise rule-like answers via IR-Triplets, (ii) focusing on SLMs for agentic settings Belcak et al. (2025), and (iii) pairing performance with spectral diagnostics.

Contributions. (i) A balanced, auditable corpus of inductive IR-Triplets spanning ten forms; (ii) a schema-constrained extraction pipeline with evidence spans and validation; (iii) a systematic SLM study (0.5B–9B) showing robust ID gains and meaningful OOD transfer; (iv) post-hoc spectral analysis and OLS disentangling scale effects from performance, motivating subspace-level probes for inductive robustness.

Paper outline. Section §2 formalizes the ten inductive forms (§2.1) and describes the triplet-extraction pipeline (§2.2) together with our (C, Q, A) supervision interface and SLM motivation. Section §3 details the experimental setup, metrics and results, including OOD transfer and OLS analyses linking performance to spectral shifts. Section §4 discusses implications for agentic systems. Section §5 concludes with limitations and future work.

2 INDUCTIVE REASONING DATA SET CONSTRUCTION

Throughout, *induction* refers to any reasoning that is non-deductive inference where conclusions are not guaranteed by the premises. This encompasses not only generalization from examples to rules, but also abductive reasoning (inference to best explanation), analogical reasoning, causal inference under uncertainty, and probabilistic updating. Unlike deductive reasoning, where conclusions follow necessarily from premises, inductive conclusions extend beyond the given information and remain tentative, subject to revision with new evidence.

2.1 FORMS OF INDUCTIVE REASONING

We operationalize induction as a family of reasoning patterns that extend beyond the given evidence forms commonly used in analysis and science. These patterns are not mutually exclusive; many instances combine multiple forms. Each dataset item is tagged with a primary (and optional secondary) form and rendered as a triplet (C, Q, A): C (evidence/context), Q (hypothesis or decision query), A (a rule-like answer, often with conditions/qualifiers).

Our work focuses on ten canonical forms of inductive reasoning:

- **Enumerative generalization:** many cases $\rightarrow$ class-wide claim.
- **Statistical generalization:** sample frequency $\rightarrow$ population rate (with error).

Table 1: Inductive forms with per-form citations and (C, Q, A) templates.

Form	Template	
Enumerative generalization Hurley (2018); Henderson (2022)	**C:**	Many A_i are P
	Q:	What about all/most A?
	A:	(Probably) all/most A are P
Statistical generalization Hurley (2018)	**C:**	In sample S, $f\%$ of A are P
	Q:	Population rate?
	A:	(Probably) $\approx f\%$ (within error)
Statistical syllogism Wesley (1971); Henry (1961)	**C:**	About $f\%$ of A are P; x is A
	Q:	What about x?
	A:	(Probably) x is P
Predictive induction Hurley (2018)	**C:**	All observed A at $t \leq T$ were P
	Q:	Next A?
	A:	(Probably) P
Analogical reasoning John (2021); Douglas et al. (2008)	**C:**	B shares relevant features F with A; A has G
	Q:	Does B have G?
	A:	(Probably) yes, c.p.
Causal induction (Mill) Mill (1974)	**C:**	When C varies, E covaries (controls held); or only difference is C
	Q:	Is C a cause of E?
	A:	(Probably) contributes to E
Abduction (IBE) Douven (2021)	**C:**	Data D observed
	Q:	Which H best explains D?
	A:	H^* (best fit/simplicity/scope/coherence) is probably true
Bayesian induction William (2022); Colin & Peter (2006)	**C:**	Prior $P(H)$; $P(D\|H) > P(D\|\neg H)$
	Q:	$P(H\|D)$?
	A:	Increases via Bayes' rule
Carnapian confirmation Rudolf (1952); Franz (2024)	**C:**	Evidence E stated in a formal language
	Q:	Degree of support for H?
	A:	Compute $c(H, E)$
Default / nonmonotonic Christian & Aldo (2019)	**C:**	Normally $A \Rightarrow P$; x is A; no defeaters known
	Q:	What to conclude now?
	A:	Tentatively x is P (retract on exception)

- **Statistical syllogism:** base rate $\rightarrow$ single-case prediction.
- **Predictive induction:** extrapolate pattern to next/unseen case.
- **Analogical reasoning:** transfer from similar source to target (ceteris paribus).
- **Causal induction (Mill):** covariation/difference suggests causal contribution.
- **Abduction (Inference to the Best Explanation (IBE)):** choose hypothesis that best explains data.
- **Bayesian induction:** update belief via Bayes' rule.
- **Carnapian confirmation:** language-relative confirmation $c(H, E)$.
- **Default/nonmonotonic:** retractable "normally" rules with explicit defeat.

Table 1 provides formal templates for each reasoning form, while Table 2 illustrates concrete examples.

In our framework, *Context* $\approx$ *Facts* (the evidential state), the *Question* targets the proposition to be resolved, and the *Answer* records the conclusion one would obtain by applying an appropriate

Table 2: Examples of inductive forms rendered as (C, Q, A).

Form	**Example** (C, Q, A)	
Enumerative generalization	**C:**	200 sampled ravens were black.
	Q:	What about ravens generally?
	A:	(Probably) ravens are black.
Statistical generalization	**C:**	In a random poll, 62% favor policy X.
	Q:	What about the electorate?
	A:	Roughly 62% (within the poll's margin of error).
Statistical syllogism	**C:**	95% of scheduled flights land safely; this is a scheduled flight.
	Q:	Will it land safely?
	A:	Probably yes.
Predictive induction	**C:**	A machine produced in-spec parts all week.
	Q:	What about the next part?
	A:	Probably in-spec.
Analogical reasoning	**C:**	Model B matches Model A on engine, weight, and aero; A gets 50 mpg.
	Q:	What is B's mpg?
	A:	Probably near 50 mpg.
Causal induction (Mill)	**C:**	Removing additive C is the only change; yield drops.
	Q:	Why the drop?
	A:	Absence of C likely caused it.
Abduction (IBE)	**C:**	Puddles, overcast sky, wet lawn.
	Q:	What best explains these facts?
	A:	It rained (better than sprinklers/hose, given context).
Bayesian induction	**C:**	A diagnostic test is reasonably sensitive and specific; result is positive.
	Q:	How likely is the disease now?
	A:	Higher than prior, computable via Bayes' rule.
Carnapian confirmation	**C:**	Observed $F(a_1), \ldots, F(a_n)$.
	Q:	How strongly is $\forall x\, F(x)$ supported?
	A:	By $c(\forall x\, F(x), E)$ under a chosen inductive method.
Default / nonmonotonic	**C:**	Birds normally fly; Tweety is a bird; no info about penguins/ostriches.
	Q:	Can Tweety fly?
	A:	Tentatively yes; retract if Tweety is a penguin.

Rule to those Facts; yet the triplet $(\textit{Context},\ \textit{Question},\ \textit{Answer})$ is strictly more general than the pair $(\{\textit{Facts}\},\ \textit{Rule})$. The pair is *mechanism-centric*; compatible with auditability, swapping rules, and checking defeaters; because it presupposes an explicit inference calculus encoded as Rule. By contrast, the triplet is *data/task-centric* and serves as a behavioral specification: it captures what was asked and what was concluded regardless of whether the internal procedure is symbolic, statistical, heuristic, or tool-augmented. This makes triplets ideal for dataset/benchmark construction and for supervising task-taking agents; they can also carry calibrated outputs (e.g., posteriors, confidence bounds, rationales) within the Answer while remaining agnostic to the internal calculus. In practice, both can co-exist; store (C, Q, A) for supervision/evaluation and $(\{\textit{Facts}\},\ \textit{Rule})$ for provenance and explainability; ensuring that the Answer is reproducible from the explicit Rule given the same Context, while the triplet remains the more portable, interoperable, and model-agnostic abstraction (indeed, $(\{\textit{Facts}\},\ \textit{Rule})$ is a special case of $(\textit{Context},\ \textit{Question},\ \textit{Answer})$).

2.2 FROM TEXTUAL NARRATIVES TO (C, Q, A) TRIPLETS

Given an input narrative x (news blurb, analyst note, research paper, decision traces from a machine learning or a data science pipeline etc.), we cast triplet extraction as schema-constrained instruction following. A frontier foundation model (e.g., ChatGPT, Claude) is prompted with (i) a *closed* output schema; (ii) the inventory of inductive reasoning forms defined in the previous section; and (iii) few-shot exemplars per form. The model must find one or more non overalpping triplets for every form $\phi \in \Phi$ (e.g., statistical generalization, causal abduction). Each triplet has the format (C, Q, A) where: C is a *minimal sufficient* evidence span lifted from x (not world knowledge), Q is a focused query answerable from C, and A is a short, atomic proposition that resolves Q given C. To curb hallucinations and improve auditability, we enforce hard constraints in the prompt (JSON keys, length limits, "no external facts" rule), require the model to return evidence offsets for C, and run automatic post-hoc checks: structural validation (schema/length), deduplication, and an entailment test that C supports A under Q via an auxiliary Natural Language Inference (NLI) verifier. We further calibrate quality by self-consistency (sampling k extractions and taking majority/median). This keeps triplets tightly grounded in the source text while aligning each (C, Q, A) instance with an explicit inductive pattern, enabling downstream training, evaluation, and error analysis. One sample from each form is included in table 3. As summarized in Table 4, the IR Triplets dataset is well balanced across the ten inductive forms ($N = 1{,}807$). The largest class, *Enumerative induction*, accounts for 11.3% of triplets, while the smallest, *Inference to the Best Explanation (abduction)*, accounts for 8.3%. The average count per form is $\bar{n} = 180.7$ triplets with a coefficient of variation of $\approx 8.9\%$, indicating modest dispersion. This balance helps limit form-specific bias and enables robust evaluation of inductive reasoning behaviors across diverse inference types.

3 INDUCTIVE REASONING DATA SET EVALUATION

3.1 EXPERIMENTAL SETUP.

We evaluate under both *in-distribution* (ID) and *out-of-distribution* (OOD) regimes. A small language model (SLM) is fine-tuned on the training split of our inductive–reasoning corpus, while the unmodified pretrained model serves as the baseline. We then assess *both* models on (i) the held-out ID subset of our corpus and (ii) a fully OOD benchmark—the DEER dataset Yang et al. (2024). This 2×2 design (model: baseline vs. fine-tuned; data: ID vs. OOD) quantifies in-distribution gains and tests whether improvements transfer to data generated by a distinct process.

Supervised fine-tuning approach. We employ parameter-efficient fine-tuning via LoRA adapters to adapt small language models to our inductive reasoning task. Our training methodology prioritizes computational efficiency while maintaining model expressiveness through targeted adaptation of attention mechanisms.

We structure training examples as contextual reasoning problems, where prompts present observational evidence followed by a focused question, with models trained to generate rule-like conclusions. This format mirrors the natural flow of inductive inference: from specific observations to general principles. The tokenization strategy masks prompt tokens during loss computation, focusing optimization exclusively on answer generation, a design choice that prevents the model from simply memorizing input patterns while encouraging principled reasoning from context.

Our LoRA configuration targets the four attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) with rank $r = 16$, balancing adaptation capacity against overfitting risk. We adopt a conservative learning rate of 2×10^{-4} with cosine scheduling and modest warmup, training for three epochs with effective batch size 16. This regime proved sufficient for convergence across our model suite while avoiding the instabilities often observed in small model fine-tuning.

For evaluation, we employ deterministic decoding with light repetition penalties to ensure consistent, comparable outputs across models.

3.2 METRICS AND SPECTRAL DIAGNOSTICS

ROUGE-L (Longest Common Subsequence). ROUGE-L Lin (2004) measures overlap between a candidate string $X = (x_1, \ldots, x_n)$ and a reference string $Y = (y_1, \ldots, y_m)$ via the length of their

Form	Context (C)	Question (Q)	Answer (A)
Enumerative induction	Bachelor's or PhD holders tend to stay, whereas master's degree holders are more likely to leave.	Lina holds a master's degree and no other standout attributes are noted. Is she likely to stay with the company?	Probably not: master's degree holders show higher attrition in the observed data.
Statistical generalization	In 120 monthly CPI observations during the 2010s, extreme outliers were rare, and seasonality explained virtually none of the variance.	What can we infer about the outlier rate for similar stable expansion periods?	It is likely very low (near zero), though sampling uncertainty means it may not be exactly zero.
Statistical syllogism	From 1980–1999, December often ranked among the highest months; overall, seasonality explained essentially none of the variance and the effects were tiny.	For a specific January near this period, should we expect CPI to sit slightly below the annual average?	Probably slightly below, but the effect is weak and unreliable given the negligible overall seasonality.
Predictive induction	The series is trend-dominated with very high lag-1 correlation; a quick forecast hinted at minor near-term drift but was low quality.	Absent new information, what should we expect for the next period?	A modest continuation of the upward trend, with appropriate caution due to model limitations.
Causal induction (Mill-style / ATE)	Change-point screening found shifts centered on early 1984, consistent with disinflation and shifting energy dynamics in that decade.	Did mid-1980s policy/energy shifts probably contribute to CPI regime changes?	Probably—timing and consistency suggest these factors increased the likelihood of structural adjustments, though causation is not certain.
Inference to the Best Explanation (abduction)	Smooth upward trend with non-stationary levels, negligible annual seasonality, moderate multi-year rhythms ($\approx$ 4–20 years), and no outliers.	Which hypothesis best explains the data: strong calendar seasonality, or drift with occasional regime shifts and multi-year rhythms?	A drift-dominated process with occasional regime shifts and moderate multi-year rhythms best explains the observed patterns.
Analogical reasoning	CPI level is non-stationary with high persistence (random-walk-like). Standard guidance recommends differencing (month-over-month changes).	By analogy to random-walk assets, should CPI modeling prioritize differences over levels?	Yes; the similarity supports modeling the differenced series rather than levels.
Bayesian induction	Annual seasonality is weak and calendar-based CPI signals have limited value; there is strong evidence of persistence and non-stationarity.	How should this evidence update a prior belief that calendar effects are profitable to trade?	Lower that belief and favor strategies that account for regimes, while not ruling out seasonality entirely.
Carnapian inductive logic (confirmation)	Classical decomposition confirms weak seasonality; FFT-based diagnostics highlight multi-year cycles rather than a strong 12-month rhythm.	Does this evidence E confirm hypothesis H that seasonality is weak in this regime?	Yes; E increases $c(H, E)$—it raises the degree of support for H.
Nonmonotonic/default (defeasible) reasoning	No extreme shocks detected; seasonality is weak; persistence is random-walk-like. A short-horizon mixed-frequency forecast was flagged as low quality due to a frequency mismatch.	What default modeling rule should practitioners follow, and what defeaters could overturn it?	**Default:** difference CPI and avoid calendar-based trades; hedge persistence. **Defeaters:** credible evidence of strong seasonality, major regime shifts, or large shocks.

Table 3: Sample IR triplets by inductive form from our dataset.

Longest Common Subsequence (LCS), denoted $\mathrm{LCS}(X, Y)$. It defines recall and precision as

$$R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X, Y)}{m}, \qquad P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X, Y)}{n},$$

Table 4: IR Triplets by Inductive Form (N = 1,807).

Form	Count	Share (%)
Enumerative induction	204	11.3%
Predictive induction	198	11.0%
Statistical syllogism	194	10.7%
Causal induction	190	10.5%
Statistical generalization	189	10.5%
Nonmonotonic/default (defeasible) reasoning	174	9.6%
Bayesian induction	173	9.6%
Carnapian inductive logic (confirmation)	171	9.5%
Analogical reasoning	164	9.1%
Inference to the Best Explanation (abduction)	150	8.3%
Total	**1,807**	**100.0%**

and combines them with an F-score (often F_1 or F_β):

$$\mathrm{ROUGE{-}L} = F_\beta = \frac{(1+\beta^2)\, P_{\mathrm{LCS}}\, R_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}}.$$

We report the average ROUGE-L across examples. Intuitively, ROUGE-L rewards long, in-order matches without requiring strict contiguity, making it well suited to judge semantic faithfulness at the phrase level in free-form generations.

Spectral tail index Martin & Mahoney (2021). Let $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ be a learned weight matrix with singular values $\{\sigma_i\}_{i=1}^{r}$ ($r = \mathrm{rank}(W)$). Empirically, the top of the spectrum in many trained networks follows a heavy-tailed law; fitting a power law to the upper tail of the eigenvalue or singular-value distribution,

$$\Pr(\lambda > x) \;\propto\; x^{-\alpha} \quad \text{for large } x,$$

yields the *spectral tail index* $\alpha > 0$. Smaller α indicates heavier tails and stronger long-range correlations (stronger implicit regularization), while larger α indicates lighter tails (more Wishart-like behavior). We estimate α by maximum-likelihood power-law fitting on the high-eigenvalue tail with standard goodness-of-fit checks. In our tables, a drop in the tail index after fine-tuning reflects heavier-tailed, more correlated spectra.

Stable rank Martin & Mahoney (2021). The *stable rank* of W is an effective dimensionality that is robust to small singular values:

$$\mathrm{srank}(W) \;=\; \frac{\|W\|_F^2}{\|W\|_2^2} \;=\; \frac{\sum_{i=1}^{r} \sigma_i^2}{\sigma_1^2}, \qquad 1 \le \mathrm{srank}(W) \le r.$$

Lower stable rank means energy is concentrated in a few leading directions (greater compression/structure), whereas higher stable rank indicates more spread across modes. The consistent decrease we observe post fine-tuning suggests stronger low-rank structure and reduced effective capacity, which often correlates with better in-distribution generalization, though it does not by itself guarantee out-of-distribution transfer.

Reporting. For all metrics we also report a *Delta* defined as the absolute change (Fine Tuned − Baseline) and, in parentheses, the relative change as a percentage of the baseline.

3.3 Results

Task performance. Across all SLMs, supervised fine-tuning (SFT) on our inductive–reasoning triplets consistently improves ROUGE-L on the in-distribution (ID) test set, with substantial absolute

gains for several smaller models (e.g., TinyLlama-1B: $+0.125$, Gemma2-2B: $+0.112$). On the out-of-distribution (OOD; DEER) benchmark, transfer is positive for most models (e.g., Gemma2-2B: $+0.088$, Qwen-1.5B: $+0.032$), and near-zero for a few (e.g., TinyLlama-1B, Yi-9B). Overall, SFT on IR triplets improves text-level fidelity without task-specific tuning.

Spectral effects. Post-hoc spectral measurements show that fine-tuning compresses model spectra: the spectral tail index drops by ≈ 30–70% and stable rank declines by ≈ 40–55% across models. In heavy-tailed random-matrix analyses, such shifts are associated with more compact, lower-rank representations, which are often linked to better generalization.

Linking performance and spectra (OLS). To relate performance gains to spectral changes while accounting for scale, we ran four OLS regressions with spectral deltas as outcomes and ΔROUGE-L and model size as predictors (ID and OOD; Table 9). Two clear findings emerge:

1. **Model size is a strong predictor of spectral compression.** Larger models exhibit significantly more negative spectral deltas (stronger compression) both ID and OOD ($p<0.02$ for stable rank; $p\leq 0.01$ for tail index).

2. **ΔROUGE-L is not a significant predictor once size is controlled.** Coefficients on ΔROUGE-L are negative but not statistically significant (all $p>0.18$). Thus, performance gains are not simply a function of how much spectral compression occurs; scaling effects dominate variance in spectral shifts.

Do spectral shifts predict ROUGE gains? We also regressed the ROUGE-L delta on spectral deltas and size (Table 10). Coefficients for the spectral terms have the expected signs—more stable-rank compression (more negative Δ) tends to associate with larger ROUGE gains—but none are statistically significant once we control for size (ID: $p=0.335$; OOD: $p=0.492$ for ΔStable Rank; tail-index $p\geq 0.69$). Size itself is not significant in this specification ($p\geq 0.64$). Overall explanatory power is modest to low (ID uncentered $R^2=0.67$; OOD $R^2=0.33$), and residual diagnostics indicate non-normality, making effect-size uncertainty substantial with $N=10$.

Taken together with the "forward" regressions (spectral deltas $\sim$ ROUGE $+$ size), these results suggest that SFT on IR triplets reliably *improves* ROUGE and *induces spectral compression*, but the *magnitude* of ROUGE gains does not linearly track the *amount* of global spectral compression after accounting for scale. This points to representation *where* compression happens (task-relevant subspaces) as a more promising explanatory factor than *how much* compression happens globally.

Table 5: ROUGE-L (In-Distribution)

Model	Baseline	Fine-Tuned	Delta
Tiny Llama 1B	0.106	0.231	$+0.125$ ($+118\%$)
Qwen 1.5B	0.111	0.178	$+0.067$ ($+60\%$)
Gemma2 2B	0.132	0.245	$+0.112$ ($+86\%$)
Gemma 7B	0.085	0.119	$+0.033$ ($+40\%$)
DeepSeek 7B	0.089	0.227	$+0.138$ ($+155\%$)
Olmo 7B	0.111	0.264	$+0.153$ ($+138\%$)
LLama 8B	0.072	0.302	$+0.23$ ($+319\%$)
Apertus 8B	0.038	0.059	$+0.021$ ($+55\%$)
Gemma2 9B	0.107	0.171	$+0.064$ ($+60\%$)
Yi 9B	0.106	0.166	$+0.06$ ($+57\%$)

Table 6: ROUGE-L (Out-of-Distribution DEER dataset)

Model	Baseline	Fine-Tuned	Delta
Tiny Llama 1B	0.049	0.049	+0.000 (+0%)
Qwen 1.5B	0.039	0.071	+0.032 (+82%)
Gemma2 2B	0.037	0.125	+0.088 (+238%)
Gemma 7B	0.034	0.037	+0.003 (+9%)
DeepSeek 7B	0.045	0.05	+0.023 (+51%)
Olmo 7B	0.071	0.073	+0.002 (+3%)
LLama 8B	0.111	0.21	+0.099 (+89%)
Apertus 8B	0.038	0.059	+0.021 (+55%)
Gemma2 9B	0.078	0.096	+0.018 (+23%)
Yi 9B	0.132	0.167	+0.035 (+26%)

Table 7: Spectral Tail Index

Model	Baseline	Fine-Tuned	Delta
Tiny Llama 1B	5.56	2.18	-3.38 (-60.8%)
Qwen 1.5B	7.73	3.08	-4.65 (-60.2%)
Gemma2 2B	3.67	1.46	-2.21 (-60.2%)
Gemma 7B	2.54	0.54	-2.00 (-78.7%)
DeepSeek 7B	9.66	5.35	-4.31 (-44.6%)
Olmo 7B	2.09	0.31	-1.78 (-85.2%)
Llama 8B	6.89	3.04	-3.85 (-55.9%)
Apertus 8B	14.63	7.57	-7.06 (-48.3%)
Gemma2 9B	3.47	1.28	-2.19 (-63.1%)
Yi 9B	5.58	2.42	-3.16 (-56.6%)

4 Discussion: IR-Triplet–tuned SLMs as the backbone of reasoning agents

Our fine-tuned SLMs have not been tested as agent reasoners; nonetheless, the IR-Triplet interface suggests a path to domain-general ampliative [2] inference that is compatible with tool use and planning. In fact, we propose a setup where the *agent does not call a general-purpose LLM*. Instead, every "reasoning" call is routed to a compact *small language model fine-tuned on* (C, Q, A) *inductive-reasoning triplets* (IR-Triplets). This matches the job specification of agentic AI, where

[2] Describes reasoning or inference where the conclusion goes beyond what is strictly contained in the premises.

Table 8: Stable Rank

Model	Baseline	Fine-Tuned	Delta
Tiny Llama 1B	131.75	67.57	-64.18 (-48.7%)
Qwen 1.5B	131.67	67.97	-63.70 (-48.4%)
Gemma2 2B	155.38	82.67	-72.71 (-46.8%)
Gemma 7B	229.19	127.78	-101.41 (-44.2%)
DeepSeek 7B	234.28	121.94	-112.34 (-48.0%)
Olmo 7B	222.71	115.30	-107.41 (-48.2%)
Llama 8B	218.68	110.72	-107.96 (-49.4%)
Apertus 8B	246.66	111.36	-135.30 (-54.9%)
Gemma2 9B	206.911	111.86	-95.05 (-45.9%)
Yi 9B	230.68	114.34	-116.34 (-50.4%)

Table 9: OLS regressions linking ROUGE-L improvements to spectral shifts, controlling for model size. Each cell reports coefficient (std. err.). N=10 models. Uncentered R^2 reported by statsmodels.

	Δ Spectral Tail Index (%)		Δ Stable Rank (%)	
Predictor	In-Dist.	OOD	In-Dist.	OOD
Δ ROUGE-L (%)	−0.104 (0.103) p=0.342	−0.071 (0.130) p=0.602	−0.116 (0.079) p=0.181	−0.089 (0.104) p=0.416
Size (B parameters)	−6.849 (2.077)** p=0.011	−7.991 (1.637)*** p=0.001	−4.699 (1.603)** p=0.019	−5.914 (1.306)*** p=0.002
R^2 (uncentered)	0.831	0.816	0.834	0.807
F-statistic (p-value)	19.69 ($<$0.001)	17.78 (0.001)	20.03 ($<$0.001)	16.69 (0.001)

Notes: Dependent variables are percentage deltas (fine-tuned minus baseline) in spectral tail index and stable rank. Coefficients for Size indicate that larger models undergo stronger spectral compression (more negative Δ), both ID and OOD. ΔROUGE-L coefficients are negative but not statistically significant once size is controlled. $^{**}p<0.05$, $^{***}p<0.01$.

Table 10: OLS with ΔROUGE-L (%) as the dependent variable and spectral deltas + size as predictors. Cells show coefficient (std. err.) with two-sided p on the next line. N=10.

	In-Distribution	Out-of-Distribution
Δ Tail Index (%)	0.939 (2.235) $p = 0.687$	0.763 (1.991) $p = 0.713$
Δ Stable Rank (%)	−2.826 (2.731) $p = 0.335$	−1.765 (2.432) $p = 0.492$
Size (B params)	4.888 (10.129) $p = 0.644$	1.286 (9.021) $p = 0.891$
R^2 (uncentered)	0.669	0.325
F-statistic (p-value)	4.72 (0.0417)	1.13 (0.402)

autonomous systems must make *ampliative* inferences under uncertainty. The triplet format serves as a behavioral specification that captures what was asked and what was concluded, without committing to brittle, symbolic internals.

This induces an *inductive control loop* that mirrors how tools are actually used: the agent (i) reads a noisy *Context* C (logs, tables, time-series traces), (ii) formulates a task-relevant *Question* Q (a subgoal or decision criterion), and (iii) queries the IR-Triplet–tuned SLM for an *Answer* A that is actionable and rule-like (e.g., a gating condition, selection policy, or hypothesis). For example, a smart-home agent might observe repeated thermostat adjustments on cloudy afternoons (C), ask whether this pattern implies a preference (Q), and infer a proactive policy for the next cloudy day (A). By training directly on (C, Q, A), the model learns to map evidence to concise rules and to expose its decision boundary in natural language, improving *decomposition* (which sub-questions to ask), *tool selection* (which diagnostic to run next), *stopping* (when sufficient evidence has accrued), and *self-checking* (justifications that external code can verify).

Replacing a frontier LLM with an IR-Triplet–tuned SLM yields *substantially lower latency and cost per decision* (smaller parameter count, lower VRAM footprint, easier edge or on-prem deployment) while preserving task performance in the intended domain. Moreover, grounding outputs in the provided C and constraining them to concise, rule-like A reduces the *surface area for hallucination*: answers are shorter, context-anchored, and amenable to programmatic checks. Because (C, Q, A) makes the reasoning step explicit and structured, proposals can be logged, linted against constraints, unit-tested on holdout data, and revised by humans which would improve controllability, traceability, and safety. The result is a smaller, faster model that lowers inference cost and *reduces hallucination risk* in practice, while fitting naturally into tool-using pipelines.

5 CONCLUSION

This work shows that fine-tuning small language models (SLMs) on a principled corpus of inductive–reasoning triplets reliably improves text-level fidelity and alters model spectra in ways consistent with capacity consolidation. Grounding data construction in well-known forms of ampliative inference (from enumerative induction to Bayesian updating) yields a compact (Context, Question, Answer) training signal that transfers across architectures. Empirically, across ten SLMs (0.5B–9B), we observe consistent *in-distribution* gains; mean ROUGE-L improvement around $\sim$60–70% with several large absolute increases (e.g., TinyLlama-1B $+0.125$; Llama-8B $+0.230$). On the *out-of-distribution* DEER benchmark, improvements are heterogeneous but frequently positive (e.g., Gemma2-2B $+0.088$, $+238\%$), indicating non-trivial transfer without task-specific tuning.

Post-hoc spectral diagnostics reveal strong compression after IR-Triplets supervised fine tuning: the spectral tail index typically declines by $\sim$45–80% (median $\sim$60%), while stable rank declines by $\sim$44–55% across models. Controlled regressions clarify how these phenomena relate. When predicting spectral deltas from ROUGE-L gains and size, *size* is a significant predictor of compression (ID and OOD), whereas the ROUGE-L gain coefficient is not. Conversely, when predicting ROUGE-L gains from spectral deltas and size, neither spectral metric nor size is significant with $N = 10$, and explanatory power is modest (especially OOD). Taken together, these results support the following picture: IR-Triplet SFT reliably improves ROUGE-L and induces measurable spectral compression, but the *magnitude* of ROUGE-L gains does not linearly track the *amount* of global compression once scale is controlled. In other words, spectral compression is a robust *correlate* of effective fine-tuning, while *where* compression occurs in representation space may matter more than *how much* compression occurs overall.

As with any early study, our analysis has limitations. Limitations include the small model set (N=10) and potential departures from ideal assumptions; future work aims at strengthening inference via more robust uncertainty quantification and sensitivity analyses to confounds. With model size already controlled, the next step is to probe representation change directly (e.g., selectivity, cross layer/task alignment, rule or class conditioned behavior) and to run targeted interventions that adjust capacity allocation (architecture, sparsity, regularization) to isolate what drives transfer in inductive reasoning. Overall, the evidence suggests that training SLMs on inductive reasoning triplets is a simple, model agnostic lever for improving generalization while reorganizing capacity into more compact and utilitarian representations; forthcoming work will test the durability and scope of these gains under stronger controls, at larger scale, and within agentic frameworks.

REFERENCES

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. 2025.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, Bing Yin, and Yizhou Sun. Inductive or deductive? rethinking the fundamental reasoning abilities ofllms. 2024.

Francois Chollet. On the measure of intelligence. 2019.

Strasser Christian and Antonelli G. Aldo. Non-monotonic logic. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2019.

Howson Colin and Urbach Peter. *Scientific Reasoning: The Bayesian Approach*. Open Court, 3rd edition, 2006.

Walton Douglas, Reed Christopher, and Macagno Fabrizio. *Argumentation Schemes*. Cambridge University Press, 2008.

Igor Douven. Abduction. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2021.

Huber Franz. Rudolf carnap. inductive logic. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2024.

Leah Henderson. The problem of induction. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2022.

Kyburg Henry. *Probability and the Logic of Rational Belief*. Wesleyan University Press, 1961.

Patrick Hurley. *A Concise Introduction to Logic*. Cengage Learning, 14th edition, 2018.

Haibo Jin, Peiyan Zhang, Man Luo, and Haohan Wang. Reasoning can hurt the inductive abilities of large language models. 2025.

Norton John. Analogy and analogical reasoning. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2021.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. 2024.

Chunyang Li, Weiqi Wang, Tianshi Zheng, and Yangqiu Song. Patterns over principles: The fragility of inductive reasoning in llms under noisy observations. *ACL*, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. 2004.

Charles Martin and Michael Mahoney. Implicit self-regularization in deep neural networks. 2021.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive*, volume 7-8 of *Collected Works of John Stuart Mill*. University of Toronto Press, 1974.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. 2023.

Carnap Rudolf. *The Continuum of Inductive Methods*. University of Chicago Press, 1952.

Fali Wang, Zhiwei Zhang, Xianre n Zhang, , and Zongyu Wu. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. 2024a.

Ruocheng Wang, Eric Zelikman1, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models. *ICLR*, 2024b.

Salmon Wesley. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, 1971.

Talbott William. Bayesian epistemology. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2022.

Tong Xiao and Jingbo Zhu. Foundations of large language models. 2025.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. 2024.