

GEAR: A GENERAL EVALUATION FRAMEWORK FOR ABDUCTIVE REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Since the advent of Large Language Models (LLMs), research has primarily focused on improving their instruction-following and deductive reasoning abilities. Yet a central question remains: can these models truly discover new knowledge, and how can we evaluate this ability? In this work, we address this gap by studying abductive reasoning—the process of generating plausible hypotheses to explain observations. We introduce **General Evaluation for Abductive Reasoning (GEAR)**, a new general-purpose, fully automated, transparent, and label-free evaluation paradigm that overcomes limitations of prior approaches. GEAR evaluates a set of hypotheses using three metrics: **consistency** (each hypothesis correctly explains the given observations), **generalizability** (consistent hypotheses make meaningful predictions on unseen inputs), and **diversity** (the set of hypotheses covers many distinct predictions and patterns). Built this way, GEAR is scalable (no human gold answers needed), reliable (transparent, deterministic scoring aligned with classical abduction), and open-ended (scores improve only when models produce new, plausible hypotheses, unlike existing static benchmarks that saturate once accuracy is high). Using GEAR, we conduct a fine-grained study of nine LLMs on four popular abduction benchmarks (1,500 problems), generating 50,340 candidate hypotheses. GEAR reveals model differences and insights that are obscured by prior gold-answer-based or purely human evaluations. We further propose a momentum-based curriculum training strategy that dynamically adjusts GEAR-derived training data by learning velocity: it begins with what the model learns faster and shifts toward harder objectives such as generating diverse hypotheses once the model is confident on foundational objectives (e.g., instruction following and consistency). Without gold-label supervision, this strategy improves all three GEAR objectives—consistency, generalizability, and diversity—and these gains transfer to established abductive-reasoning benchmarks. Taken together, GEAR provides a principled framework that not only evaluates abduction but also supplies label-free, scalable training signals that help LLMs produce more diverse and reliable hypotheses. We will release code and data upon acceptance.

1 INTRODUCTION

In the current AI community, there are many competing definitions of abductive reasoning. The most widely adopted one is Harman’s view of abduction as *inference to the best explanation* (IBE) (Harman, 1965; Douven, 2021). Although this definition is simple and intuitive, it suffers from key limitations when applied to real-world settings, making benchmarks and evaluations built on it problematic.

First, IBE does not specify what counts as “best,” and the criteria vary across contexts. In some cases, simplicity is prioritized; in others, novelty or explanatory power is preferred. As a result, IBE-based benchmarks often select a single “gold” hypothesis according to annotators’ subjective judgments, yielding heterogeneous and unreliable labels (Zhao et al., 2023; Okasha, 2000; Cabrera, 2023). **Second**, multiple plausible hypotheses typically coexist. Real-world observations can often be explained in several ways, depending on how the data are conceptualized. This limitation appears across popular abductive benchmarks such as MINI-ARC (Kim et al., 2022), ACRE (Zhang et al.,

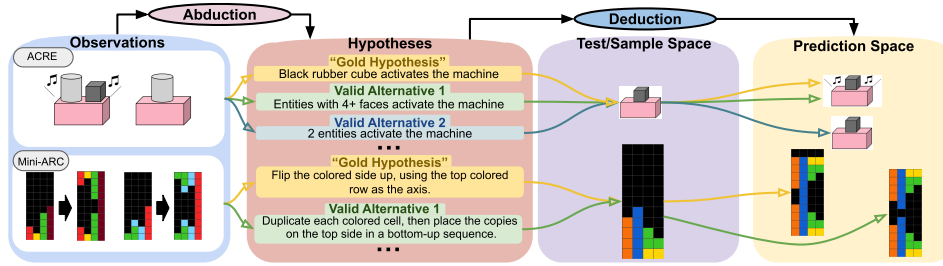


Figure 1: Underdetermination in abductive reasoning: a logically sound, observation-consistent hypothesis need not coincide with the annotated “gold” hypothesis; both can fit the seen data yet disagree on held-out predictions. Single-gold labeling can mask the plurality of valid explanations.

2021), LIST FUNCTIONS (Rule, 2020), and ARC-2025 (Chollet, 2019). For example, in Figure 1, the annotated “gold” hypothesis and other valid alternatives may all be well supported by sound abductive reasoning. However, existing evaluation usually excludes these alternatives by enforcing agreement with a single annotated “gold” hypothesis under a vague “best” standard. Philosophical accounts of abduction emphasize that hypotheses are often *evidentially underdetermined*: given finite data and background assumptions, more than one hypothesis may be well supported (Quine, 2014; Stanford, 2023), the ability to generate other plausible, well-supported hypotheses should also count as abductive success. Scientific progress relies on maintaining a diverse pool of feasible hypotheses. Although many will ultimately be falsified, they guide the design of experiments to discriminate, verify, or refute competing explanations, thereby enriching our knowledge. Even in current mature fields, credible alternative hypotheses often remain.

To address these limitations, we use Peirce’s original definition of abduction, which frames it as a more general task of generating hypotheses from given observations (Frankfurt, 1958; Burks, 1946; Minnameier, 2004; Peirce, 1974). Building on this foundation, we propose GEAR, a new framework for systematically evaluating abductive reasoning. GEAR is grounded in three classical criteria for good scientific hypotheses: (1) **Consistency**. A hypothesis must not contradict observed facts, ensuring compatibility with existing evidence. (2) **Generalizability**. A good hypothesis extends beyond the observed data by making testable predictions on unseen cases. In Popper’s terms, better hypotheses carry higher empirical content: they make riskier, more precise claims and thus invite more opportunities for refutation (Popper, 2005; 2014). We operationalize generalizability as the coverage of unseen inputs on which a hypothesis yields determinate predictions; larger coverage indicates greater generality. Accordingly, when two hypotheses fit the observed data, we prefer the broader (more falsifiable) one; if it withstands more severe tests, it is more strongly corroborated and more robust. (3) **Diversity**. A hypothesis should contribute a genuinely new perspective rather than echo existing ones, to avoid premature convergence on a single explanation. In the spirit of Chamberlin’s multiple working hypotheses and Platt’s strong inference (Chamberlin, 1965; Platt, 1964), we favor *sets of hypotheses* that articulate competing mechanisms testable by critical evidence. We quantify diversity with two complementary measures: γ -diversity, the average number of unique predictions per input across the hypotheses set (set-level variety), and β -diversity, the dissimilarity of prediction patterns between hypothesis pairs (dispersion). Higher diversity indicates broader causal coverage and multiple viewpoints on the observations; points of disagreement highlight decisive experiments and help accelerate scientific progress.

With GEAR, we re-examine four popularly used abduction benchmarks MINI-ARC, ACRE, LIST FUNCTIONS, and ARC-2025 across nine LLMs—spanning API-access models (GPT-o1 (OpenAI, 2024), GPT-4.1-mini (OpenAI, 2025a), GPT-o4-mini (OpenAI, 2025b)) and open-source models (LLAMA-3.3-70B, LLAMA-3.1-8B (Grattafiori et al., 2024), QWEN-2.5-72B, QWEN-2.5-7B (Qwen et al., 2025), GEMMA-2-9B (Team et al., 2024), NEXTCODER-7B (Aggarwal* et al., 2025))—and we find that— (1), Consistency remains hard, 70B-class models produce only 20% consistent hypotheses; (2) Consistency shows no significant correlation with the size of the initial observation set; (3), Model size is weakly related to abductive diversity—larger models do not necessarily generate more diverse hypotheses; and (4) existing gold-answer evaluations overlook the underdetermination inherent to abduction, around 80% of equally plausible hypotheses are labeled incorrect, and even the “accepted” hypotheses can differ substantially. Unlike prior frameworks that depend on gold answers or human raters, GEAR is label-free and fully automated, yielding dense, scalable signals that directly train models to generate consistent, generalizable, and diverse

hypothesis sets. We convert these signals into optimization targets for preference-based RL and fine-tune base models with LoRA, so that improvements on GEAR’s objectives are optimized end-to-end without gold supervision. To stabilize learning and broaden coverage across objectives, we introduce a momentum-based curriculum learning strategy that dynamically adjusts GEAR-derived training data by learning velocity: training begins with fast-to-learn, foundational objectives (instruction following and consistency) and shifts toward harder reasoning objectives that foster diverse hypothesis generation as competence increases. This procedure raises GEAR scores and transfers as accuracy gains on standard abductive-reasoning benchmarks across multiple model families (e.g., Qwen-2.5-7B, Llama-3.1-8B, NextCoder-7B).

2 RELATED WORK

Reasoning: evolution from non-defeasible to defeasible. Non-defeasible (deductive) reasoning preserves truth under added premises, whereas defeasible reasoning allows conclusions to be revised when new evidence appears (Yu et al., 2024). Abduction belongs to the latter: following Peirce, abduction proposes candidate hypotheses for observed facts, and deduction derives precise, testable predictions from a hypothesis (Frankfurt, 1958; Burks, 1946; Minnameier, 2004; Peirce, 1974). A key contrast with deduction is abduction’s reliance on broad background knowledge (commonsense and domain-specific), which naturally yields multiple distinct yet plausible hypotheses for the same observation (He & Chen, 2025). Early symbolic systems struggled here due to narrow knowledge bases, whereas LLMs pretrained on large corpora make such abductive tasks more tractable (Yang et al., 2023; He & Chen, 2025). Despite its central role in discovery, abduction remains understudied compared with the extensive focus on deduction in AI (Niu et al., 2024; Liu et al., 2025; Yu et al., 2024; Huang & Chang, 2023; He & Chen, 2025).

Gold- and human-based evaluation for abductive reasoning. Current practice largely relies on two strands. *Gold answer-based* evaluation compares model outputs to a single reference either (i) at the *hypothesis level* using BLEU/ROUGE or embedding metrics such as BERTScore (Yang et al., 2024a; Qi et al., 2024; Movva et al., 2025; Bowen et al., 2024; Hua et al., 2025; Young et al., 2022), or (ii) at the *behavior level* by matching input–prediction pairs implied by the reference hypothesis (Sinha et al., 2019; Weston et al., 2015; Balepur et al., 2024; Shi et al., 2023; Wang et al., 2024; Rule, 2020; Chollet, 2019; Liu et al., 2024; Li et al., 2025; Chen et al., 2025; He et al., 2025). Despite scalability, single-reference matching is ill-suited to abduction: it rejects many plausible, logically sound alternatives that merely differ from the annotated answer; it is also costly (expert labeling) and unstable due to non-monotonic judgments and low inter-annotator agreement (Young et al., 2022). Complementarily, *human evaluation* is often used for qualities that are hard to algorithmically quantify (e.g., novelty, excitement) (Zhao et al., 2024; Qi et al., 2024; Yang et al., 2024b; Hu et al., 2024; Yang et al., 2025), but it is expensive, hard to reproduce or scale, and inherently subjective—particularly acute for abduction, where outcomes depend on rater expertise, instructions, and context, and small samples limit statistical power. In sum, both strands conflict with the essence of abductive reasoning: instead of testing agreement with a single “gold” explanation or subjective impressions, evaluations should assess a model’s capacity to propose *multiple*, novel, and plausible explanatory hypotheses when underlying causes are unknown.

3 GEAR

We introduce the **General Evaluation for Abductive Reasoning** (GEAR), an evaluation paradigm that scores hypotheses using *reference-free, transparent criteria* rather than agreement with a single gold answer. Unlike existing benchmarks that evaluate only a single generated hypothesis at a time, GEAR evaluates a *set* of hypotheses along three dimensions. An LLM exhibits stronger abductive proficiency when it can produce a hypothesis set that (i) correctly explains the given observations (**Consistency**), (ii) yields meaningful predictions on unseen inputs (**Generalizability**), and (iii) offers non-redundant alternatives rather than superficial variants (**Diversity**). Basic notation appears in Table 1.

Table 1: Abductive reasoning begins with a set of observations $\mathbb{O} = \{o_1, o_2, \dots\}$, where each $o_i := (\text{in}_i, \text{out}_i)$ is an input–output pair (e.g., $o_1 = (\text{floor}, \text{wet})$, $o_2 = (\text{air}, \text{humid})$). A hypothesis “it rained” can be represented as a function f_{rain} mapping inputs to outputs, e.g., $f_{\text{rain}}(\text{floor}) = \text{wet}$, $f_{\text{rain}}(\text{sky}) = \text{cloudy}$. Each hypothesis f has an input domain $\mathbb{D}$; outside this domain, predictions may be undefined or uninformative (e.g., $f_{\text{rain}}(\text{Math})$).

Symbol	Meaning
$\mathbb{O} = \{o_1, o_2, \dots\}$	Set of observations to be explained
$o_i := (\text{in}_i, \text{out}_i)$	Observation as an input–output pair
f	Hypothesis function
$\mathbb{F}$	A set of hypotheses
$\mathbb{D}$	(Effective) input domain of hypothesis f
M	A set-size measure (e.g., cardinality $ \cdot $)
$M(\mathbb{D})$	Size of the input domain $\mathbb{D}$
$\tilde{f}$	Trivial hypothesis that memorizes all seen observations
$\mathbb{S} = \{\text{in}_1, \text{in}_2, \dots\}$	Problem-specific sample space of candidate inputs
$\mathbb{P}_f := \{(\text{in}, f(\text{in})) : \text{in} \in \mathbb{S}\}$	Prediction space of f on $\mathbb{S}$

3.1 CONSISTENCY

Consistency is the most fundamental requirement of a hypothesis: it must not conflict with observed facts. Formally, a generated hypothesis f is consistent with the observation set $\mathbb{O}$ if $\forall (\text{in}_i, \text{out}_i) \in \mathbb{O}, f(\text{in}_i) = \text{out}_i$. This criterion guarantees agreement with all known observations and underlies most existing evaluations of hypothesis generation, including gold answer–based evaluations.

3.2 GENERALIZABILITY

Given several consistent hypotheses, a more general hypothesis is one that yields predictions on a broader set of unseen cases. A more general hypothesis confers two advantages: (1) it can be applied in more future situations, increasing its practical utility; and (2) because it makes predictions for more situations, it can be tested more extensively and—if it survives—becomes correspondingly more robust (Popper, 2005; 2014). Formally, for two hypotheses f_1 and f_2 with respective input domains $\mathbb{D}_1$ and $\mathbb{D}_2$ and a set-size measure M , if $M(\mathbb{D}_1) > M(\mathbb{D}_2)$, then f_1 is considered more general than f_2 , with a simple example.

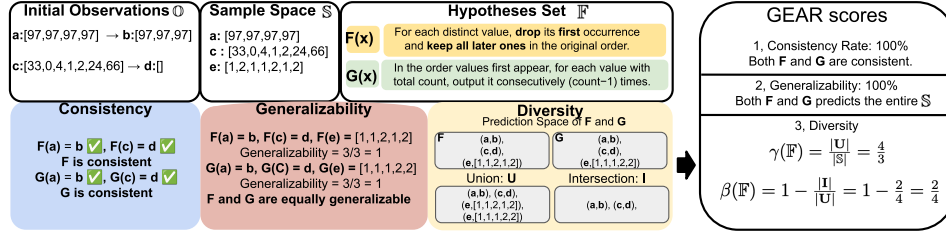
For example, given three observations $o_1 = (1, 1)$, $o_2 = (10, 1)$, and $o_3 = (100, 1)$, a trivial lookup-table hypothesis $\tilde{f}$ that merely memorizes these pairs is consistent yet fails to generalize. In contrast, $f_1(n) = \text{rev}(n)$ (digit-reversal on integers), $f_2(x) = x/x$ for $x \neq 0$ (undefined at $x = 0$), and the constant hypothesis $f_3(x) \equiv 1$ are all consistent but differ in generalizability. Under the simple size measure $M(\mathbb{D}) = |\mathbb{D}|$, we have: $\{1, 10, 100\} \subset \mathbb{N}_0 \subset \mathbb{R} \setminus \{0\} \subset \mathbb{R} \Rightarrow \tilde{f} \prec f_1 \prec f_2 \prec f_3$.

The effective domain $\mathbb{D}$ and its size measurement M are not static but depend on the problem context and representation. For instance, in arithmetic, f_1 (reversal) is defined on integers, whereas in programming tasks the same hypothesis applies to strings, lists, and other finite sequences, making f_1 more general than f_2 in that setting. Since it is generally infeasible to determine a global domain $\mathbb{D}$ across all conceivable contexts, GEAR evaluates generalizability relative to a pre-defined *problem-specific sample space* $\mathbb{S}$ shared across all hypotheses under comparison, together with its associated measurement M . This sample space serves as the operational domain for evaluation and provides the basis for comparing the generalizability of different hypotheses.

3.3 DIVERSITY

Prior work typically measures hypothesis diversity with black-box text similarity metrics such as BERTScore (Zhang et al., 2019)—which capture surface-level semantic overlap rather than underlying explanatory mechanisms—or with subjective, annotator-based judgments. In contrast, following the Multiple Working Hypotheses view (Chamberlin, 1965; Platt, 1964), we define diversity directly from *prediction patterns*: distinct hypotheses should be *separable* by some input in $\mathbb{S}$. Concretely, for two consistent hypotheses f_1, f_2 and an unseen input $\text{in}_x \in \mathbb{S}$, if $f_1(\text{in}_x) \neq f_2(\text{in}_x)$, then f_1 and f_2 are separable on in_x ; the more such inputs exist, the more diverse the hypotheses are.

In light of classical ecological diversity theory—specifically β - and γ -diversity (Whittaker, 1960)—we adapt set-based diversity ideas to hypothesis sets, not as a one-to-one import from ecol-

Figure 2: GEAR with a live example with $|\mathbb{F}| = 2$, $|\mathbb{S}| = 3$.

ogy but as a structural analogy over prediction patterns. Concretely, we define γ as the average number of unique predictions per input over a set of hypotheses $\mathbb{F}$ and β as the mean pairwise Jaccard dissimilarity between prediction sets on a shared sample space.

γ -diversity (average unique predictions per input). Let $\mathbb{P}_f = \{(\text{in}, f(\text{in})) : \text{in} \in \mathbb{S}\}$. Define

$$\gamma(\mathbb{F}; \mathbb{S}) := \frac{1}{M_1(\mathbb{S})} M_2\left(\bigcup_{f \in \mathbb{F}} \mathbb{P}_f\right) = \frac{1}{|\mathbb{S}|} \sum_{\text{in} \in \mathbb{S}} \left| \{(\text{in}, f(\text{in})) : f \in \mathbb{F}\} \right|,$$

Under cardinality $M_1 = M_2 = |\cdot|$ and $|\mathbb{F}| = m$, $\gamma \in [1, m]$. A hypothesis proposer that generates near-duplicate hypotheses will yield $\gamma \approx 1$ because most hypotheses agree on almost all inputs. Conversely, a proposer that views the observations from diverse perspectives and produces genuinely novel hypotheses will achieve a larger γ , approaching $\gamma \approx m$ when predictions are mutually distinct for every input. Notably, M_1 is the size measure on the sample (input) space, whereas M_2 is the size measure on the prediction space; the two need not be identical.

β -diversity (prediction-pattern dispersion). Measure pairwise dispersion via the Jaccard dissimilarity between prediction sets, and then average across all pairs:

$$d_J(f_i, f_j) := 1 - \frac{M(\mathbb{P}_{f_i} \cap \mathbb{P}_{f_j})}{M(\mathbb{P}_{f_i} \cup \mathbb{P}_{f_j})}, \quad \beta(\mathbb{F}; \mathbb{S}) := \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq m} d_J(f_i, f_j).$$

Example. Let $\mathbb{S} = \{0, 1, 2\}$, $\mathbb{P}_{f_1} = \{(0, 1), (1, 2), (2, 3)\}$, $\mathbb{P}_{f_2} = \{(0, 1), (1, 2), (2, 2)\}$, and $M = |\cdot|$. *Generalizability:* for each f , $G(f) = |\mathbb{P}_f|/|\mathbb{S}|$. Thus $G(f_1) = G(f_2) = 3/3 = 1$. *Set coverage:* $\bigcup_i \mathbb{P}_{f_i} = \{(0, 1), (1, 2), (2, 2), (2, 3)\}$, so $\gamma = |\bigcup_i \mathbb{P}_{f_i}|/|\mathbb{S}| = 4/3$. *Diversity:* the intersection size is 2 and the union size is 4, hence the Jaccard distance $d_J(f_1, f_2) = 1 - \frac{2}{4} = \frac{1}{2}$; with a single pair, $\beta = \frac{1}{2}$. (If fewer than two consistent hypotheses are generated, we set $\beta = 0$.) An additional example from LIST FUNCTIONS is shown in Fig. 2.

While γ and β are mathematically related, neither uniquely determines the other. The formal relationship and proofs are provided in Appendix F.

Because Consistency, Generalizability, and Diversity have precise mathematical definitions, they can be computed directly—without human labor or auxiliary black-box models—at evaluation time. Consequently, GEAR is (1) *scalable*: it runs automatically on any newly generated hypotheses; (2) *reliable*: all metrics are transparently defined and computed, including Diversity, which GEAR measures by underlying mechanisms rather than via black-box proxies; and (3) *open-ended*: GEAR evaluates how many genuinely different explanatory perspectives a model can produce and, because the hypothesis space is in principle unbounded, it imposes no upper limit on valid hypotheses.

4 LLM EVALUATION SETTINGS

Benchmarks. We use four widely used abductive benchmarks: MINI-ARC (Kim et al., 2022), ACRE (Zhang et al., 2021), LIST FUNCTIONS (Rule, 2020), and ARC-2025 (Chollet, 2019). These gold-answer benchmarks split observations into $\mathbb{O}_{\text{train}}$ and $\mathbb{O}_{\text{test}}$. Unlike the traditional setting—where models see only $\mathbb{O}_{\text{train}}$ and are judged on $\mathbb{O}_{\text{test}}$ —GEAR measures *how many* consistent hypotheses a model can produce and *how diverse* they are. Because several datasets provide only a few observations per problem (e.g., 2–3 train and 1 test), we pool all observations into $\mathbb{O}_{\text{all}} := \mathbb{O}_{\text{train}} \cup \mathbb{O}_{\text{test}}$ to enable broader analyses. We choose these datasets because (i) hypotheses are expressible in formal languages (e.g., executable programs), allowing deterministic evaluation

without extra adaptation; and (ii) each dataset provides sufficiently diverse observations to seed abduction.

Sampling initial observations. We study the effect of observation size by using $n \in \{1, 2, 3, 4\}$ observations. For ACRE, outputs are Boolean (`on/off`); with a single observation, label semantics may be ambiguous, so we use $n \in \{2, 3, 4\}$ and ensure at least one `on` and one `off`. When n observations are required, we sample n pairs from $\mathbb{O}_{\text{all}}$ without replacement to form $\mathbb{O}_n$. We randomly select 100 problems per dataset. For three datasets we use $n = 4$ (total $3 \times 100 \times 4 = 1200$), and for ACRE we use $n = 3$ (total $1 \times 100 \times 3 = 300$), yielding 1,500 problems overall.

Generation protocol. Given $\mathbb{O}_n$, we prompt with a dataset-agnostic template P_{init} to obtain the first hypothesis f_1 . We then iterate with P_{iter} , which lists previously generated hypotheses $\mathbb{F}_{t-1} = \{f_1, \dots, f_{t-1}\}$ and requests a *new* f_t that is (i) consistent with $\mathbb{O}_n$ and (ii) distinct from all $f \in \mathbb{F}_{t-1}$. Both P_{init} and P_{iter} are shared across datasets (see P_{init} and P_{iter} in the Appendix G).

Stopping rule. Since it is infeasible to enumerate all hypotheses an LLM could generate for a given problem, we use a *quality-triggered* early stop: generation for a problem halts once the model emits three “bad” hypotheses—unparseable, inconsistent with $\mathbb{O}_n$, or non-novel w.r.t. $\mathbb{F}_{t-1}$. This keeps runs finite and comparable while allowing stronger models to produce more consistent, novel hypotheses before stopping. Formal criteria and thresholds are in Appendix B.

Sample space $\mathbb{S}$. We assess generalizability and diversity on a fixed, dataset-specific $\mathbb{S}$. Since exhaustive coverage is infeasible, we construct $\mathbb{S}$ via simple, reproducible rules with a fixed seed. For LIST FUNCTIONS and ACRE, whose input domains are explicit (lists; entity triples), we stratify by input size: include the empty case, include all singletons, and uniformly subsample the remainder. For MINI-ARC/ARC-2025, naive enumeration of visual grids mostly yields noise; thus $\mathbb{S}$ is the set of unique input grids from the official splits after canonicalization and deduplication. Full construction details and cardinalities are in Appendix A.

5 ANALYSIS OF LLM EVALUATIONS

5.1 MAIN ANALYSIS: LLM PERFORMANCE UNDER GEAR

Across nine LLMs we collect 50,340 hypotheses, with 17,835 *consistent* and 32,505 *inconsistent*; of the latter, 4,346 fail format/parseability (e.g., non-executable) and 28,159 contradict at least one observation (Fig. 3).

Overall volume and consistency. Under the quality-triggered early stop, the hypotheses-per-problem reflect overall generation capacity. GPT-04-MINI and GPT-01 produce the most consistent hypotheses and the highest consistency rates, yielding ~ 2 –4 more hypotheses per problem than the next tier (panel (a)) and correspondingly higher consistency (panel (c)).

Effect of initial observation size. Increasing I/O pairs from 1 to 4 tightens constraints: both β and γ diversity decline (panels (d)–(e)), while the consistency rate remains comparatively stable (panel (c)). Early stopping is increasingly triggered by the novelty threshold rather than inconsistency.

Instruction following. Most models adhere to the required format (panel (b)); LLAMA-3.1-8B is an outlier, often appending free-form text or violating the code template, harming parseability. (§6 shows RL substantially closes this gap.)

Model size vs. abductive diversity. Among open-source models, parameter count correlates weakly with abductive diversity: LLAMA-3.3-70B trails GEMMA-2-9B and QWEN-2.5-72B on β -diversity (panel (d)), and QWEN-2.5-72B is only on par with these smaller models—suggesting data and training procedure may matter more than raw size for abductive reasoning.

Generalizability. Because $\mathbb{S}$ pertains to each dataset’s distribution, consistent hypotheses achieve high coverage ($\approx 95\%$ on average; panel (f)). Corner cases remain in MINI-ARC/ARC-2025 (e.g., infinite loops). Since GPT-01 and GPT-04-MINI contribute a large share of consistent hypotheses on these harder datasets, their macro-averaged coverage appears lower—not due to intrinsically weaker generality, but because the solved problems are more challenging.

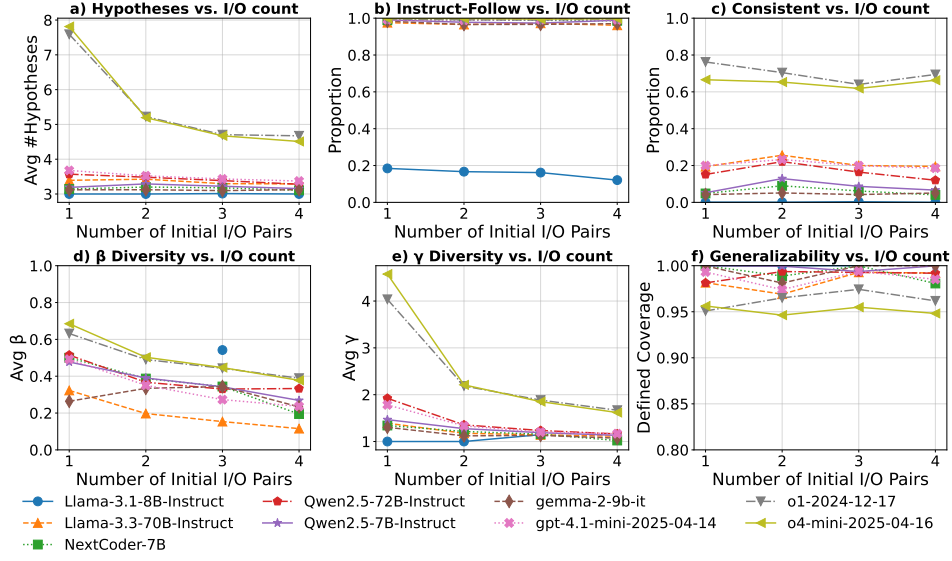


Figure 3: Model performance under GEAR. Metrics are *macro-averaged per problem*: compute the value per problem, then take the unweighted mean. Panels (a)–(c) use *all* generated hypotheses; panels (d)–(f) restrict to *consistent* hypotheses, since diversity and generalizability are only meaningful for consistent sets.

5.2 SIMULATION STUDY 1: ABDUCTIVE REASONING IS DEFEASIBLE

As noted in § 2, abduction is *defeasible*: a hypothesis generated by logical and sound abductive reasoning need not coincide with the “gold answer.” Benchmarks that enforce a single gold hypothesis thus reward label matching rather than generating alternative, plausible explanations; and we also show that without extensive test cases previous benchmarks *cannot* even explicitly distinguish an alternative hypothesis from the gold-labeled one as intended.

Empirical illustration. We previously generated 17,835 hypotheses consistent with initial observations $\mathbb{O}_n$ ($n \in \{1, 2, 3, 4\}$, sampled from $\mathbb{O}_{\text{all}}$). For each problem, fixing $\mathbb{O}_n$ and its consistent hypotheses set $\mathbb{F}$, we sample $m \in \{1, 2, 3, 4\}$ hidden test pairs $\mathbb{O}_m \subset \mathbb{O}_{\text{all}} \setminus \mathbb{O}_n$ and repeat each (n, m) five times. A hypothesis *passes* if it agrees with all test cases in $\mathbb{O}_m$. We report (a) pass rate and (b) γ -diversity of survivors, averaged over problems (Fig. 4a–b).

Pass rates decline as m increases but do not collapse; even at $m=4$, a nontrivial fraction remains valid. Survivors also retain diversity—for example, on ACRE at $m=4$ the passed set has $\gamma \approx 1.5$, i.e., roughly 1.5 distinct predictions per input on the shared sample space. Because all candidates are LLM-generated, this is a conservative lower bound on plausible explanations. As available observations and background knowledge broaden, underdetermination grows, and single-gold-answer metrics increasingly misclassify reasonable alternatives.

5.3 SIMULATION STUDY 2: GEAR IS GENERAL

In this section, we show that GEAR is **general** rather than task-specific. Across varied abductive datasets, higher GEAR scores predict a greater chance that the set contains a hypothesis explaining unseen (hidden) observations, thereby aligning with existing evaluations of good abduction and indicating cross-task applicability.

For each problem, we sample $m \in \{1, 2, 3, 4\}$ hidden cases $\mathbb{O}_m \subset \mathbb{O}_{\text{all}} \setminus \mathbb{O}_n$ and draw a context $\mathbb{F}_c$ of size $c \in \{0, 1, 2\}$ from hypotheses consistent with $\mathbb{O}_n$; we discard any context with a member already passing $\mathbb{O}_m$. Let $\mathbb{F}_{\text{consistent}}$ be the full set of hypotheses consistent with $\mathbb{O}_n$ and from the remaining consistent pool $\mathbb{F}_{\text{consistent}} \setminus \mathbb{F}_c$, we enumerate unordered pairs (f_a, f_b) and, for each $f \in \{f_a, f_b\}$, compute marginal diversity gain for $\rho \in \{\gamma, \beta\}$ as $\Delta_\rho(f) = \rho(\mathbb{F}_c \cup \{f\}; \mathbb{S}) - \rho(\mathbb{F}_c; \mathbb{S})$, define generalizability as $g(f) = \frac{|\mathbb{F}_f|}{|\mathbb{S}|}$, the coverage of f over the sample space $\mathbb{S}$, and an average GEAR score: $\text{Score}(f) = \frac{1}{3}(g(f) + \Delta_\gamma(f) + \Delta_\beta(f))$. We label $(f_{\text{chosen}}, f_{\text{rejected}})$ so that $\text{Score}(f_{\text{chosen}}) > \text{Score}(f_{\text{rejected}})$. We then test both on $\mathbb{O}_m$, declare a pass when a hypothesis explains all hidden cases, and aggregate over problems/contexts/pairs to report the odds ratio $\text{OR}_m = \Pr(\text{pass} \mid f_{\text{chosen}}) / \Pr(\text{pass} \mid f_{\text{rejected}})$.

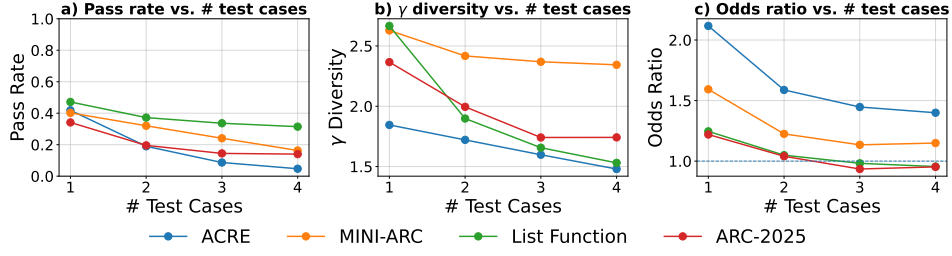


Figure 4: Simulation Study Results

Figure 4(c) shows that OR_m tends toward 1 as m increases, yet in most settings $OR_m > 1$: the higher-GEAR-score candidate is more likely to pass the hidden cases, aligning with the idea that a more diverse prediction space on $\mathbb{S}$ increases the chance of covering the unseen evidence $\mathbb{O}_m$.

6 REINFORCEMENT LEARNING WITH GEAR

6.1 TRAINING DATA PREPARATION

Building on the simulation study above, we posit that learning *better abduction*—operationalized as generating a hypothesis set that aligns better with GEAR—which naturally improves the chance that at least one hypothesis fits unseen observations and thereby *generalizes* to downstream tasks, without requiring any gold hypotheses as supervision.

Training data are constructed from the previously generated 50,340 hypotheses (§5.1). Within each dataset (100 problems), we sample 50 for training, 10 for validation, and hold out 40 for final evaluation. Similar to the preference-pair construction in § 5.3, for each training problem, let $\mathbb{F}_{\text{all}}$ be the set of hypotheses generated by nine LLMs (including both inconsistent and consistent hypotheses), and let $\mathbb{F}_c \subseteq \mathbb{F}$ be a small hypothesis context of size $c \in \{0, 1, 2\}$ sampled from $\mathbb{F}_c \subseteq \mathbb{F}_{\text{consistent}}$. We then enumerate all unordered hypothesis pairs (f_a, f_b) from $\mathbb{F}_{\text{all}} \setminus \mathbb{F}_c$. For each $f \in \{f_a, f_b\}$, we assign a preference in three stages: (1) *instruction following / format compliance* (prefer parsable over non-parsable outputs; 159,981 pairs); (2) *consistency* (among parsable candidates, prefer those consistent with $\mathbb{O}_n$; 429,980 pairs); (3) when both are parsable and consistent, GEAR score, where we score each candidate by the same score function $\text{Score}(f)$ in § 5.3. We prefer the candidate with the higher $\text{Score}(f)$ (249,561 pairs). In total, this yields 839,522 training preference pairs. We fine-tune the base models with Direct Preference Optimization (DPO) using LoRA (rank 128, $\alpha = 256$) under 4-bit quantization with bfloat16 compute. For DPO, we randomly sample 51,200 pairs with a fixed 1:1:1 ratio across Parsing, Consistency, and GEAR preferences.

6.2 MOMENTUM-BASED CURRICULUM LEARNING

During fine-tuning we observed that the *mixture* of preference data materially affects outcomes, and different base models favor different mixtures (some benefit from earlier emphasis on format/consistency, others from earlier diversity). This led us to a momentum-based curriculum method. The intuition is to learn what improves *fastest* and is *easiest* first, then shift weight toward harder signals. We did an ablation study on the effect of each preference category (See in Appendix E). This adaptively reweights based on measured learning progress, avoiding hand-tuned fixed ratios and letting each base model gravitate to its preferred mixture.

Adaptive reweighting preferences. Unlike existing curriculum learning methods that either prescribe a static training curriculum—thereby overlooking model-specific competence differences (Bengio et al., 2009; Wang et al., 2019)—or adopt dynamic schemes that operate at the sample level (Zhou et al., 2021; Jiang et al., 2015; Sow et al., 2025), which are computationally expensive and often misaligned with the higher-level objective of producing diverse hypotheses, our approach is simple yet efficient: it dynamically delivers a skill-level curriculum—i.e., a goal-aligned schedule over preference types (Parsing/Format, Consistency, and GEAR/diversity) rather than over individual instances—aligned with GEAR objectives.

For each preference type r , keep an Exponential Weighted Moving Average (EWMA) of its probe loss and convert recent improvement into sampling weight:

$$E_r^{(t)} = (1 - \alpha) E_r^{(t-1)} + \alpha L_r^{(t)}, \quad m_r^{(t)} = \text{clip}(E_r^{(t-1)} - E_r^{(t)}, 0, m_{\text{max}}), \quad p_r^{(t)} \propto \varepsilon + m_r^{(t)}$$

Table 2: Aggregated results across nine settings.
GEAR *performance improvement*

	Instruction Following Rate	Consistency	Generalizability	β	γ
Llama-3.1-8b	0.149	0.002	1.000	0.000	1.000
Llama-3.1-8b-dpo-fixed-ratio	0.968	0.014	1.000	0.000	1.000
Llama-3.1-8b-dpo-momentum-curriculum	0.970	0.020	1.000	0.196	1.179
qwen-2.5-7b	0.988	0.037	0.987	0.032	1.001
qwen-2.5-7b-dpo-fixed-ratio	0.996	0.042	0.987	0.044	1.046
qwen-2.5-7b-dpo-momentum-curriculum	0.997	0.041	1.000	0.073	1.052
nextcoder-7b	0.863	0.013	1.000	0.140	1.081
nextcoder-7b-dpo-fixed-ratio	0.965	0.028	1.000	0.078	1.066
nextcoder-7b-momentum-curriculum	0.991	0.030	0.991	0.182	1.151
<i>Cross-task generalization with GEAR</i>					
	Avg Train Pass Rate	Avg Test Pass Rate	Top-1 Acc.	Top-2 Acc.	Top-3 Acc.
Llama-3.1-8b	0.011	0.009	0.000	0.003	0.004
Llama-3.1-8b-dpo-fixed-ratio	0.096	0.076	0.028	0.043	0.049
Llama-3.1-8b-dpo-momentum-curriculum	0.129	0.101	0.035	0.052	0.065
qwen-2.5-7b	0.198	0.154	0.035	0.059	0.068
qwen-2.5-7b-dpo-fixed-ratio	0.205	0.160	0.054	0.064	0.077
qwen-2.5-7b-dpo-momentum-curriculum	0.203	0.163	0.066	0.083	0.095
nextcoder-7b	0.151	0.121	0.025	0.044	0.055
nextcoder-7b-dpo-fixed-ratio	0.173	0.136	0.051	0.065	0.074
nextcoder-7b-momentum-curriculum	0.189	0.149	0.065	0.080	0.092

Where $r \in \mathcal{R}$ indexes preference types; $L_r^{(t)}$ is the probe loss on the validation subset at epoch t ; $E_r^{(t)}$ is its EWMA; $\alpha \in (0, 1)$ is the smoothing factor; $m_r^{(t)}$ is the clipped improvement (capped by $m_{\max} > 0$); $p_r^{(t)}$ is the per-type sampling weight (normalized across r); $\varepsilon > 0$ prevents zero weight; and $p_r^{(t)}$ is then normalized across r and clipped to $[w_{\min}, w_{\max}]$.

Experimental settings. We use $\alpha = 0.1$, $\varepsilon = 0.1$, $m_{\max} = 0.03$, $w_{\min} = 0.8$, $w_{\max} = 1.2$, and update the sampling weights every 1,280 training examples. The same hyperparameters and schedule were used for all three fine-tuned models (no per-model tuning).

For evaluation, we use the original (non-sampled) splits with $\mathbb{O}_{\text{train}}$ and $\mathbb{O}_{\text{test}}$ from held-out problems, asking each model to generate *three* hypotheses per problem. As shown in Table 2, models trained from these GEAR-derived preferences—despite never seeing held out problems—achieve higher diversity scores (β, γ) and improved Top-1/2/3 (T3) accuracies, with the momentum-based curriculum consistently outperforming the fixed-ratio baseline across the reported settings.

7 CONCLUSION & DISCUSSION

We present GEAR, a general evaluation framework for abduction that scores hypotheses by consistency, generalizability, and diversity. Across nine LLMs on four abduction benchmarks, GEAR reveals differences that gold- or human-centric evaluations miss; simulation study confirms abduction is defeasible and shows that more diverse hypothesis sets are more likely to predict hidden observations. We convert GEAR into label-free training signals and propose a momentum-based DPO curriculum that adapts preference weights with learning progress, improving hypothesis diversity and downstream accuracy without gold supervision.

Although our experiments use programmable domains, GEAR is not limited to them. The framework hinges on four ingredients: (i) a size measure M , (ii) a sample space $\mathbb{S}$, (iii) a deduction/execution oracle, and (iv) a semantic equivalence predicate. In natural-language (NL) settings these remain the same but grow harder: M should capture semantic coverage (e.g., topical/diversity weightings) rather than raw cardinality; $\mathbb{S}$ can be unlabeled yet must include many diverse probes to reveal prediction patterns; the deduction step becomes model- or tool-mediated and thus stochastic, mitigated by calibrated decoding, tool use, and repeated sampling; and equivalence must be judged semantically or via canonicalization to executable meaning representations to curb polysemy. The primary bottleneck is therefore foundational NL tooling for reliable execution and equivalence under ambiguity. GEAR’s applicability to NL tasks scales with the quality of these primitives; as they improve, the framework transfers with minimal changes—largely a swap of stronger semantic metrics and oracles.

ETHICS STATEMENT

We affirm adherence to the ICLR Code of Ethics. This work evaluates and trains language models on publicly available benchmarks (MINI-ARC, ACRE, LIST FUNCTIONS, ARC-2025) and model APIs; it does not involve human subjects, personally identifiable information, or sensitive attributes. To minimize potential harms, we restrict experiments to benign, programmable tasks. Dataset licenses and usage follow their original terms; no proprietary or private data are redistributed. There are no known conflicts of interest or third-party sponsorships that influenced results; any such relationships will be disclosed upon de-anonymization. Code and data used for evaluation will be released to facilitate auditing and responsible reuse.

REPRODUCIBILITY STATEMENT

We aim for full reproducibility. Datasets, problem sampling, and the construction of the evaluation sample spaces S are defined by simple, deterministic rules with fixed seeds; prompts for hypothesis generation (initial/iterative) are provided; and all scoring criteria (Consistency, Generalizability, β/γ Diversity) are formally specified. Training details (preference construction, DPO with LoRA, quantization, and the momentum-based curriculum schedule with hyperparameters) are described alongside exact settings. We will release code, prompts, and configuration files enabling end-to-end replication—from hypothesis generation to metric computation and fine-tuning—together with random seeds and logs upon acceptance.

REFERENCES

- Tushar Aggarwal*, Swayam Singh*, Abhijeet Awasthi, Aditya Kanade, and Nagarajan Natarajan. Nextcoder: Robust adaptation of code lms to diverse code edits. In *ICML 2025*, July 2025. URL <https://www.microsoft.com/en-us/research/publication/nextcoder-robust-adaptation-of-code-lms-to-diverse-code-edits/>.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.555. URL <https://aclanthology.org/2024.acl-long.555/>.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- Chen Bowen, Rune Sætre, and Yusuke Miyao. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 323–339, 2024.
- Arthur W. Burks. Peirce’s theory of abduction. *Philosophy of Science*, 13(4):301–306, 1946. ISSN 00318248, 1539767X. URL <http://www.jstor.org/stable/185210>.
- Frank Cabrera. Inference to the best explanation: An overview. *Handbook of abductive cognition*, pp. 1863–1896, 2023.
- Thomas C Chamberlin. The method of multiple working hypotheses: With this method the dangers of parental affection for a favorite theory can be circumvented. *Science*, 148(3671):754–759, 1965.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=6z4YKr0GK6>.

- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Igor Douven. Abduction. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.
- Harry G. Frankfurt. Peirce’s notion of abduction. *The Journal of Philosophy*, 55(14):593–597, 1958. ISSN 0022362X. URL <http://www.jstor.org/stable/2021966>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Conguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le,

Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhao- duo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

Gilbert H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965. ISSN 00318108, 15581470. URL <http://www.jstor.org/stable/2183532>.

Kaiyu He and Zhiyu Chen. From reasoning to learning: A survey on hypothesis discovery and rule learning with large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=d7W38UzUg0>. Survey Certification.

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Chen. IDEA: Enhancing the rule learning ability of large language model agent through induction, deduction, and abduction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13563–13597, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.698. URL <https://aclanthology.org/2025.findings-acl.698/>.

- Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.
- Wenyue Hua, Tyler Wong, Sun Fei, Liangming Pan, Adam Jardine, and William Yang Wang. Inductionbench: Llms fail in the simplest complexity class. *arXiv preprint arXiv:2502.15823*, 2025.
- Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL <https://aclanthology.org/2023.findings-acl.67/>.
- Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Subin Kim, Prin Phunayaphibarn, Donghyun Ahn, and Sundong Kim. Playgrounds for abstraction and reasoning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022.
- Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Mirage: Evaluating and explaining inductive reasoning process in language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 73944–73969, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b782a3462ee9d566291cfff148333ea9b-Paper-Conference.pdf.
- Emmy Liu, Graham Neubig, and Jacob Andreas. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=nUNbjMDBWC>.
- Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey, 2025. URL <https://arxiv.org/abs/2502.09100>.
- Gerhard Minnameier. Peirce-suit of truth –why inference to the best explanation and abduction ought not to be confused. *Erkenntnis*, 60(1):75–105, 2004. doi: 10.1023/B:ERKE.0000005162.52052.7f. URL <https://doi.org/10.1023/B:ERKE.0000005162.52052.7f>.
- Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=4R0pugRyN5>.
- Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*, 2024.
- Samir Okasha. Van Fraassen’s critique of inference to the best explanation. *Studies in History and Philosophy of Science Part A*, 31(4):691–710, 2000.
- OpenAI. Introducing openai o1-preview. OpenAI Blog, September 2024. URL <https://openai.com/index/introducing-openai-o1-preview/>. Accessed: 2025-09-25.
- OpenAI. Introducing gpt-4.1 in the api. OpenAI Blog, April 2025a. URL <https://openai.com/index/gpt-4-1/>. Covers GPT-4.1 mini; Accessed: 2025-09-25.
- OpenAI. Introducing openai o3 and o4-mini. OpenAI Blog, April 2025b. URL <https://openai.com/index/introducing-o3-and-o4-mini/>. Accessed: 2025-09-25.
- Charles Sanders Peirce. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press, 1974.
- John R Platt. Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *science*, 146(3642):347–353, 1964.

- Karl Popper. *The logic of scientific discovery*. Routledge, 2005.
- Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 2014.
- Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=q36rpGlG9X>.
- Willard V Quine. On empirically equivalent systems of the world. In *The Nature of Scientific Theory*, pp. 353–368. Routledge, 2014.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Joshua Stewart Rule. *The child as hacker: building more human-like models of learning*. PhD thesis, Massachusetts Institute of Technology, 2020.
- Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. Language models can improve event prediction by few-shot abductive reasoning. *Advances in Neural Information Processing Systems*, 36:29532–29557, 2023.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL <https://aclanthology.org/D19-1458/>.
- Daouda Sow, Herbert Woisetschl ger, Saikiran Bulusu, Shiqiang Wang, Hans-Arno Jacobsen, and Yingbin Liang. Dynamic loss-based sample reweighting for improved large language model pre-training. *arXiv preprint arXiv:2502.06733*, 2025.
- Kyle Stanford. Underdetermination of Scientific Theory. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhu-patiraju, L onard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram , Johan Fer-ret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Char-line Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchi-son, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Wein-berger, Dimple Vijaykumar, Dominika Rogo ni ska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kar-tikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin G rner, Mat Velloso, Mateo Wirth, Matt Davidow,

- Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Rostrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Faret, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Fuchun Wang, Xian Zhou, Wenpeng Hu, Zhunchen Luo, Wei Luo, and Xiaoying Bai. Llm assists hypothesis generation and testing for deliberative questions. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 424–436. Springer, 2024.
- Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5017–5026, 2019.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.
- Robert Harding Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338, 1960.
- Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. Logical reasoning over natural language as knowledge representation: A survey. *arXiv preprint arXiv:2303.12023*, 2023.
- Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–225, St. Julian’s, Malta, March 2024a. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-long.13/>.
- Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13545–13565, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL <https://aclanthology.org/2024.findings-acl.804/>.
- Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. MOOSE-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=X9OfMNNepI>.
- Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. AbductionRules: Training transformers to explain unexpected inputs. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 218–227, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.19. URL <https://aclanthology.org/2022.findings-acl.19/>.

- Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39, 2024.
- Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. Acre: Abstract causal reasoning beyond covariation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10643–10653, 2021.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. Abductive commonsense reasoning exploiting mutually exclusive explanations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14883–14896, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.831. URL <https://aclanthology.org/2023.acl-long.831/>.
- Wenting Zhao, Justin Chiu, Jena Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Li, and Alane Suhr. UNcommonsense reasoning: Abductive reasoning about uncommon situations. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8487–8505, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.469. URL <https://aclanthology.org/2024.naacl-long.469/>.
- Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Curriculum learning by optimizing learning dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 433–441. PMLR, 2021.

A SAMPLE SPACE $\mathbb{S}$

The sample space $\mathbb{S}$ is central to GEAR: a broader $\mathbb{S}$ exposes more inputs on which prediction patterns can be compared, thus providing a sharper lens for assessing generalizability and diversity. In practice, however, $\mathbb{S}$ must balance breadth with computational cost and with the effort required to construct valid, meaningful inputs. We therefore define $\mathbb{S}$ per dataset using simple, reproducible rules and a fixed random seed. And for simplicity, we use cardinality ($|\cdot|$) as set-size measurement in our evaluation.

LIST FUNCTIONS: In the original dataset, Inputs are lists of integers with element domain $\{0, \dots, 99\}$ and length $k \in \{0, \dots, 15\}$. After expansion, the full combinatorial space has size $\sum_{k=0}^{15} 100^k$, which is infeasible to exhaust at evaluation time. We adopt *stratified sampling by length*: (i) include the unique empty list for $k=0$; (ii) include all 100 singletons for $k=1$; (iii) for each $k \in \{2, \dots, 15\}$, uniformly sample (without replacement) up to 1,000 lists from the 100^k possibilities. This yields $|\mathbb{S}_{\text{LISTFUNC}}| = 1 + 100 + 14 \times 1,000 = 14,101$. (*One could broaden the input domain beyond the original dataset—for example by allowing negative or floating-point values, or even arbitrary real numbers—but we retain the original integer domain; in our experiments this range is already sufficient to probe diverse hypothesis behaviors.*)

ACRE: In the original dataset, each primitive entity is a triple $\langle \text{color}, \text{shape}, \text{material} \rangle$ with $\text{color} \in \{\text{blue}, \text{brown}, \text{cyan}, \text{gray}, \text{green}, \text{purple}, \text{red}, \text{yellow}\}$ (8), $\text{shape} \in \{\text{cube}, \text{cylinder}, \text{sphere}\}$ (3), $\text{material} \in \{\text{metal}, \text{rubber}\}$ (2), so the vocabulary has $8 \times 3 \times 2 = 48$ distinct entity types. An input is an *ordered* list (repetitions allowed) of c entities with $c \in \{0, \dots, 8\}$ since in the original dataset at most 8 entities are seen in one observation. We again use stratified sampling by c : include the empty list for $c=0$; include all 48 singletons for $c=1$; for each $c \in \{2, \dots, 7\}$, uniformly sample up to 1,000 lists without replacement. This gives $|\mathbb{S}_{\text{ACRE}}| = 1 + 48 + 7 \times 1,000 = 7,049$. (*Broader stress tests are possible by extending the vocabulary with new colors, shapes, or materials, but we restrict ourselves to the original schema here; empirically this space is already rich enough to discriminate among hypotheses.*)

MINI-ARC and ARC-2025: Unlike the previous two datasets, these benchmarks encode inputs as small integer grids (values in a fixed palette) that carry *visual semantics*. Naively enumerating grids within a size range produces overwhelmingly non-meaningful noise and is therefore counter-productive for abductive reasoning. Instead, we define $\mathbb{S}$ as the set of all *unique* input grids already present in the official splits (train/validation/test), after canonical serialization and deduplication of exact matches. This pragmatic choice keeps inputs semantically meaningful while avoiding sampling artifacts. It yields $|\mathbb{S}_{\text{MINIARC}}| = 767$, $|\mathbb{S}_{\text{ARC2025}}| = 4,826$.

Reproducibility: All sampling steps use a fixed random seed and uniform sampling without replacement within each stratum; we publish the materialized $\mathbb{S}$ for each dataset to ensure exact reproducibility of scores.

B STOPPING RULE AND “BAD” HYPOTHESES.

Enumerating all potential hypotheses a model could generate is infeasible. To keep generation finite and comparable across models, we adopt a *quality-triggered* early-stopping rule: generation for a problem stops once the model accumulates three “bad” hypotheses. A hypothesis is *bad* if it satisfies any of the following: (i) it cannot be parsed into executable code (format or syntax error); (ii) it is inconsistent with $\mathbb{O}_n$ (violates at least one given observation); or (iii) it lacks novelty relative to prior hypotheses. For (iii), let the shared sample space be $\mathbb{S}$ and define:

$$\text{cov}(f \mid \mathbb{F}_{t-1}) := \frac{1}{|\mathbb{S}|} \left| \{ \text{in} \in \mathbb{S} : \exists g \in \mathbb{F}_{t-1} \text{ s.t. } g(\text{in}) = f(\text{in}) \} \right|$$

We mark f as non-novel if $\text{cov}(f \mid \mathbb{F}_{t-1}) \geq 0.8$ (i.e., at least 80% of its predictions on $\mathbb{S}$ are duplicate relative to $\mathbb{F}_{t-1}$). Compared with a hard quota on the number of hypotheses, this early-stopping rule (i) prevents unbounded generation and degenerate cycling; (ii) allows stronger models to produce more consistent and novel hypotheses before stopping; and (iii) reduces the chance that a fixed quota artificially limits performance.

C TABLES

Table 3: Per-dataset GEAR scores of nine LLMs (Instruction Following, Consistency, γ , β , Generalizability).

Model	Dataset	IF	Cons.	Norm- γ (q=0)	β -Struct	Gen.
o4-mini-2025-04-16	Avg	0.9972	0.6708	2.3330	0.5036	0.9398
	ACRE	1.0000	0.9771	1.6718	0.4774	0.9816
	MINI-ARC	0.9958	0.4489	2.1715	0.4850	0.9472
	List Fns	0.9978	0.9449	3.4245	0.4850	0.9727
	ARC-2025	0.9951	0.3121	2.0644	0.5670	0.8577
gpt-4.1-mini-2025-04-14	Avg	0.9911	0.2263	1.3320	0.3457	0.9824
	ACRE	0.9926	0.5298	1.2754	0.3402	0.9863
	MINI-ARC	0.9877	0.0622	1.3708	0.3710	1.0000
	List Fns	0.9917	0.2827	1.4464	0.3001	0.9859
	ARC-2025	0.9925	0.0304	1.2355	0.3716	0.9575
o1-2024-12-17	Avg	0.9924	0.7165	2.2677	0.4892	0.9538
	ACRE	0.9888	0.9679	1.6813	0.4702	0.9998
	MINI-ARC	0.9953	0.5984	1.9939	0.4367	0.9737
	List Fns	0.9964	0.9542	3.3586	0.4962	0.9740
	ARC-2025	0.9892	0.3456	2.0372	0.5536	0.8676
Llama-3.3-70B-Instruct	Avg	0.9660	0.2469	1.1862	0.2245	0.9645
	ACRE	0.9332	0.5930	1.1729	0.2008	0.9854
	MINI-ARC	0.9772	0.0569	1.1977	0.2352	1.0000
	List Fns	0.9695	0.3095	1.2279	0.1622	0.9858
	ARC-2025	0.9842	0.0282	1.1460	0.2997	0.8867
Qwen2.5-72B-Instruct	Avg	0.9990	0.1848	1.3988	0.3996	0.9865
	ACRE	0.9987	0.4770	1.3119	0.3847	0.9998
	MINI-ARC	0.9983	0.0319	1.5144	0.5793	0.9784
	List Fns	1.0000	0.2086	1.4646	0.3511	0.9798
	ARC-2025	0.9992	0.0215	1.3042	0.2835	0.9881
gemma-2-9b-it	Avg	0.9718	0.0522	1.0715	0.1857	0.9960
	ACRE	0.9893	0.1272	1.1661	0.4065	0.9999
	MINI-ARC	0.9892	0.0063	1.0000	0.0000	1.0000
	List Fns	0.9912	0.0729	1.1218	0.1507	0.9862
	ARC-2025	0.9175	0.0025	0.9980	–	0.9980
Qwen2.5-7B-Instruct	Avg	0.9807	0.1036	1.1801	0.2714	0.9992
	ACRE	0.9597	0.3149	1.2475	0.3737	0.9971
	MINI-ARC	0.9885	0.0107	1.2061	0.3835	1.0000
	List Fns	0.9854	0.0861	1.2667	0.3285	0.9997
	ARC-2025	0.9892	0.0029	0.9999	0.0000	0.9999
Llama-3.1-8B-Instruct	Avg	0.1505	0.0020	1.0735	0.5421	1.0000
	ACRE	0.0578	0.0048	1.1469	0.5421	1.0000
	MINI-ARC	0.2217	0.0000	–	–	–
	List Fns	0.1917	0.0031	1.0000	–	1.0000
	ARC-2025	0.1308	0.0000	–	–	–
NextCoder-7B	Avg	0.9891	0.0704	1.1351	0.4410	0.9930
	ACRE	0.9911	0.2097	1.1562	0.3397	0.9999
	MINI-ARC	0.9842	0.0100	1.1550	0.5363	1.0000
	List Fns	0.9952	0.0612	1.2293	0.4470	0.9721
	ARC-2025	0.9858	0.0006	1.0000	–	1.0000

Table 4: Ablation results: GEAR score & T3 accuracy under different training settings.

GEAR Scores						
	Instruction Following Rate	Consistency	Generalizability	β	γ	
Llama-3.1-8b	0.149	0.002	1.000	0.000	1.000	
Llama-3.1-8b-dpo-fixed-ratio	0.968	0.014	1.000	0.000	1.000	
Llama-3.1-8b-dpo-momentum-curriculum	0.970	0.020	1.000	0.196	1.179	
Llama-3.1-8b-parsing-ablation	0.941	0.009	1.000	0.000	1.000	
Llama-3.1-8b-consistent-ablation	0.322	0.009	1.000	0.014	1.011	
Llama-3.1-8b-GEAR-ablation	0.258	0.005	1.000	0.000	1.000	
qwen-2.5-7b	0.988	0.037	0.987	0.032	1.001	
qwen-2.5-7b-dpo-fixed-ratio	0.996	0.042	0.987	0.044	1.046	
qwen-2.5-7b-dpo-momentum-curriculum	0.997	0.041	1.000	0.073	1.052	
qwen-2.5-7b-parsing-ablation	0.996	0.032	1.000	0.014	1.008	
qwen-2.5-7b-consistent-ablation	0.988	0.047	0.975	0.015	1.010	
qwen-2.5-7b-GEAR-ablation	0.981	0.034	1.000	0.023	1.012	
nextcoder-7b	0.863	0.013	1.000	0.140	1.081	
nextcoder-7b-dpo-fixed-ratio	0.965	0.028	1.000	0.078	1.066	
nextcoder-7b-momentum-curriculum	0.991	0.030	0.991	0.182	1.151	
nextcoder-7b-parsing-ablation	0.992	0.020	1.000	0.222	1.140	
nextcoder-7b-consistent-ablation	0.945	0.034	1.000	0.054	1.040	
nextcoder-7b-GEAR-ablation	0.980	0.015	0.976	0.202	1.145	
T3 accuracies						
	Avg Train Pass Rate	Avg Test Pass Rate	Top-1 Acc.	Top-2 Acc.	Top-3 Acc.	
Llama-3.1-8b	0.011	0.009	0.000	0.003	0.004	
Llama-3.1-8b-dpo-fixed-ratio	0.096	0.076	0.028	0.043	0.049	
Llama-3.1-8b-dpo-momentum-curriculum	0.129	0.101	0.035	0.052	0.065	
Llama-3.1-8b-parsing-ablation	0.130	0.099	0.026	0.031	0.041	
Llama-3.1-8b-consistent-ablation	0.084	0.066	0.018	0.022	0.028	
Llama-3.1-8b-GEAR-ablation	0.039	0.030	0.011	0.016	0.020	
qwen-2.5-7b	0.198	0.154	0.035	0.059	0.068	
qwen-2.5-7b-dpo-fixed-ratio	0.205	0.160	0.054	0.064	0.077	
qwen-2.5-7b-dpo-momentum-curriculum	0.203	0.163	0.066	0.083	0.095	
qwen-2.5-7b-parsing-ablation	0.198	0.154	0.045	0.055	0.066	
qwen-2.5-7b-consistent-ablation	0.224	0.176	0.059	0.075	0.086	
qwen-2.5-7b-GEAR-ablation	0.183	0.141	0.039	0.056	0.065	
nextcoder-7b	0.151	0.121	0.025	0.044	0.055	
nextcoder-7b-dpo-fixed-ratio	0.173	0.136	0.051	0.065	0.074	
nextcoder-7b-momentum-curriculum	0.189	0.149	0.065	0.080	0.092	
nextcoder-7b-parsing-ablation	0.162	0.125	0.045	0.055	0.062	
nextcoder-7b-consistent-ablation	0.200	0.153	0.060	0.074	0.080	
nextcoder-7b-GEAR-ablation	0.150	0.118	0.033	0.043	0.049	

D FIGURES

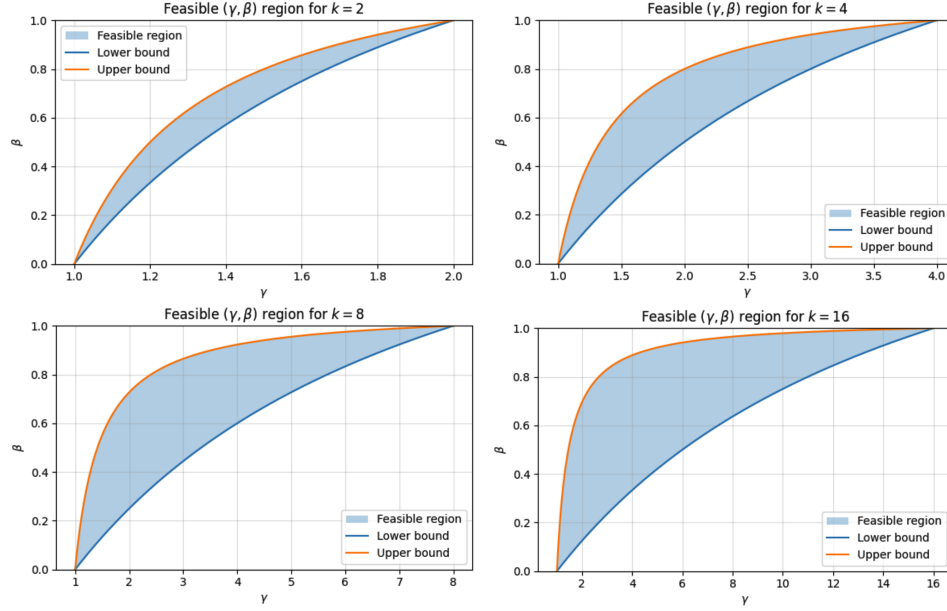
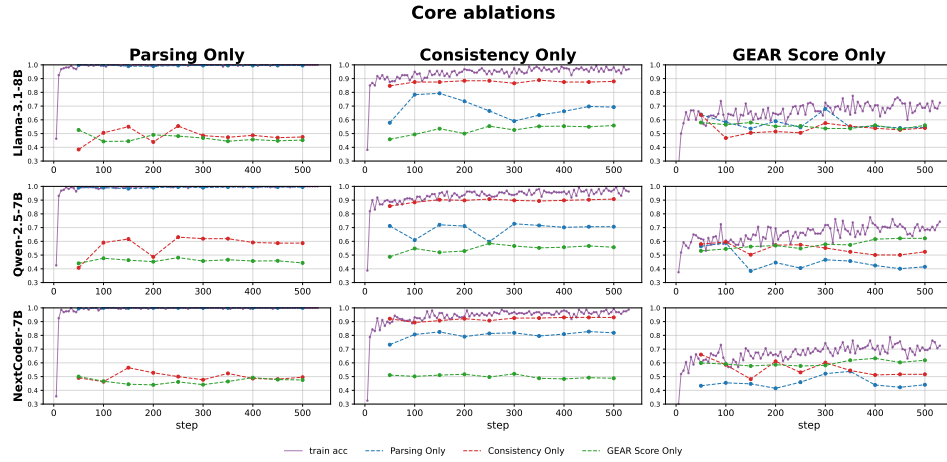
Figure 5: Feasible region for (γ, β) 

Figure 6: Training log on single-preference DPO

Fixed-ratio vs. Momentum

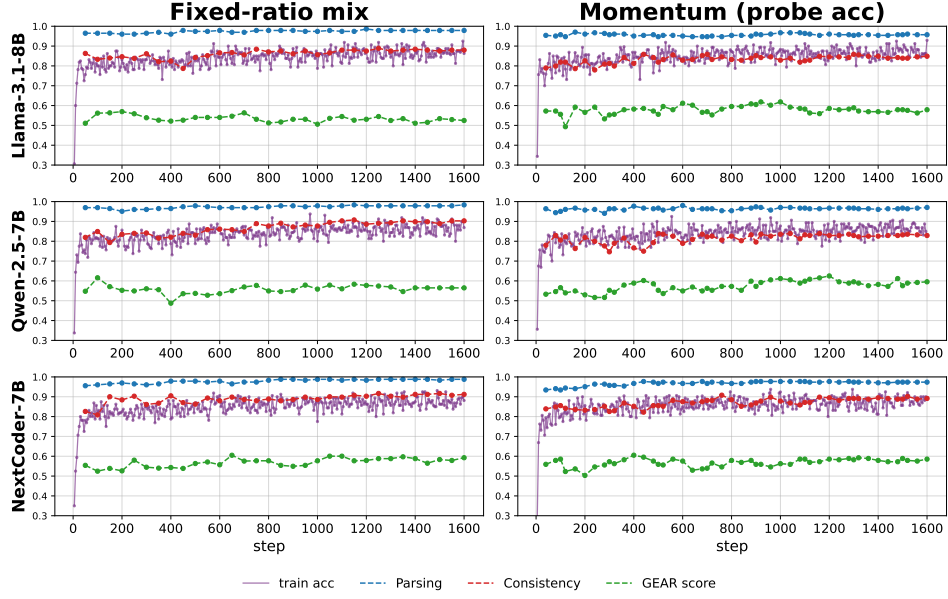


Figure 7: Training log on multi-preference DPO

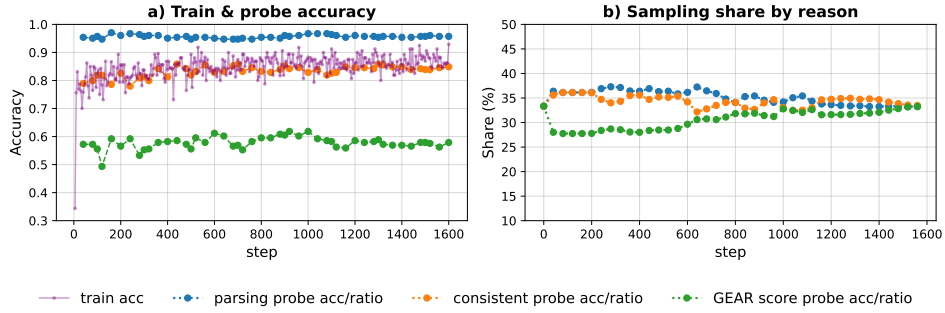


Figure 8: Momentum curriculum training log for Llama-3.1-8B

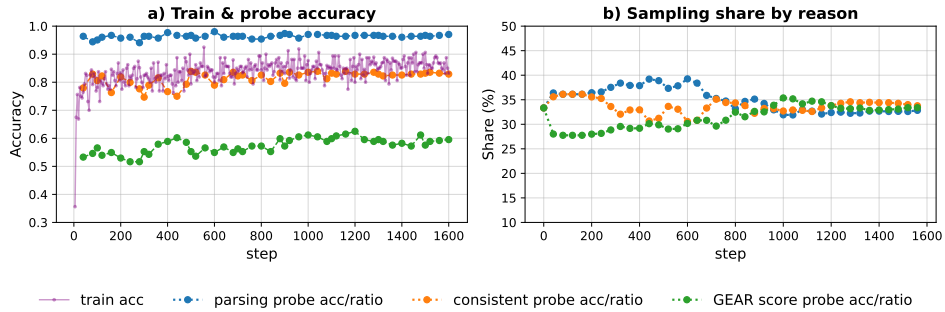


Figure 9: Momentum curriculum training log for Qwen-2.5-7B

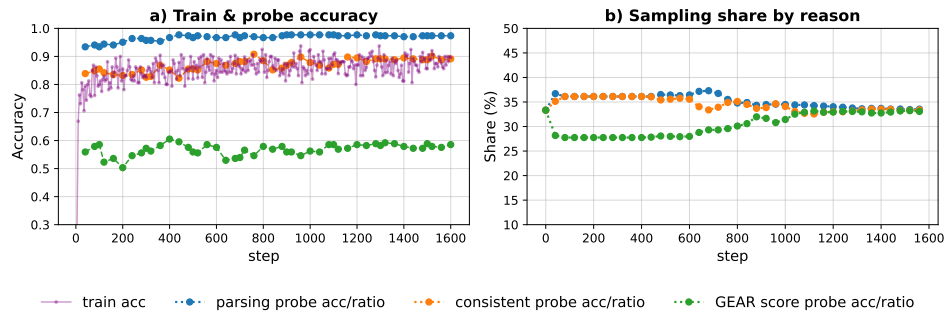


Figure 10: Momentum curriculum training log for NextCoder-7B

E TRAINING LOG ANALYSIS

Across our training signals, the three preferences exhibit a clear *hierarchy*: a hypothesis must first be parsable to be eligible for consistency checking; only hypotheses that pass consistency check can contribute to generalizability and diversity. This structure implies that optimizing for a single preference in isolation is insufficient. As shown in Figure 6, training on only one preference can raise the corresponding validation accuracy, but typically fails to improve—and often degrades—the other two. For example, parsing-only training does not translate into higher consistency or GEAR gains. Consistent with Table 4, parsing-only runs also do not yield a meaningful increase in the diversity of generated hypotheses nor in downstream T-3 accuracy. Taken together, these results indicate that leveraging all three preferences during training is necessary to realize balanced improvements across GEAR.

Figure 7 further compares fixed-ratio mixing with our *momentum curriculum*. Although momentum-trained models show slightly lower peak accuracy on parsing and consistency early on, they make steadier progress on the GEAR preference and reach roughly 60% validation accuracy during training. This translates into the strongest diversity scores and the best downstream T-3 accuracy in subsequent evaluations (see Table 4). The curriculum’s adaptive reweighting prioritizes what is learning fastest while gradually shifting emphasis to harder signals, yielding a more stable training trajectory and superior overall abduction quality.

F MATHEMATICAL RELATIONSHIP BETWEEN β & γ

Setup. Let $\mathbb{S}$ be the (finite) sample space of inputs with $|\mathbb{S}| = n$, and let $\mathbb{F} = \{f_1, \dots, f_k\}$ be k hypotheses. For each $f \in \mathbb{F}$, recall the *prediction set* on $\mathbb{S}$:

$$\mathbb{P}_f := \{(\text{in}, f(\text{in})) : \text{in} \in \mathbb{S}\}$$

We work under the equal-size assumption (standard in our experiments): every $f \in \mathbb{F}$ produces exactly one prediction per input in $\mathbb{S}$, hence $|\mathbb{P}_f| = n$.¹ Define the union size

$$U := \left| \bigcup_{f \in \mathbb{F}} \mathbb{P}_f \right|, \quad \gamma(\mathbb{F}; \mathbb{S}) = \frac{U}{|\mathbb{S}|} = \frac{U}{n} \in [1, k],$$

so γ is the average number of unique predictions per input.

Multiplicity c_x and the identity $\sum_x c_x = kn$. For each element x in the union $\bigcup_{f \in \mathbb{F}} \mathbb{P}_f$, define its *reuse multiplicity*

$$c_x := \#\{f \in \mathbb{F} : x \in \mathbb{P}_f\} \in \{1, \dots, k\}.$$

Intuitively, c_x counts in how many hypotheses’ prediction sets the *same prediction pair* $x = (\text{in}, \text{out})$ appears. By *double counting*,

$$\sum_x c_x = \sum_x \sum_{f \in \mathbb{F}} \mathbf{1}\{x \in \mathbb{P}_f\} = \sum_{f \in \mathbb{F}} \sum_x \mathbf{1}\{x \in \mathbb{P}_f\} = \sum_{f \in \mathbb{F}} |\mathbb{P}_f| = kn.$$

The quantity

$$r := \frac{kn}{U} = \frac{k}{\gamma} \in [1, k]$$

is the *average reuse multiplicity*: on average, each distinct prediction pair in the union is reused by r hypotheses.

Pairwise intersections and the definition of t . For a pair (f_i, f_j) with $i < j$, define

$$t_{ij} := |\mathbb{P}_{f_i} \cap \mathbb{P}_{f_j}| \in [0, n],$$

namely, the number of inputs in $\mathbb{S}$ on which f_i and f_j make the *same prediction pair* $(\text{in}, f(\text{in}))$. Let the total and the average pairwise intersection sizes be

$$I := \sum_{1 \leq i < j \leq k} t_{ij} = \sum_x \binom{c_x}{2} = \frac{1}{2} \left(\sum_x c_x^2 - kn \right), \quad \bar{t} := \frac{I}{\binom{k}{2}}.$$

¹If a hypothesis can be undefined on some inputs, one may attach a sentinel output $\perp$; this preserves $|\mathbb{P}_f| = n$ without changing the extremal structure below.

Jaccard similarity $\text{sim}(t)$ and why $t/(2n - t)$. For two finite sets A, B , the Jaccard *similarity* is

$$\text{Jacc}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Here $A = \mathbb{P}_{f_i}$ and $B = \mathbb{P}_{f_j}$ satisfy $|A| = |B| = n$ by assumption, so if we denote $t = |A \cap B| = t_{ij}$, then $|A \cup B| = |A| + |B| - |A \cap B| = 2n - t$, and

$$\text{sim}(t) := \text{Jacc}(\mathbb{P}_{f_i}, \mathbb{P}_{f_j}) = \frac{t}{2n - t}, \quad t \in [0, n].$$

Our β -diversity is the average Jaccard *dissimilarity* across pairs, i.e.,

$$\beta(\mathbb{F}; \mathbb{S}) = \frac{2}{k(k-1)} \sum_{i < j} (1 - \text{sim}(t_{ij})), \quad \text{so} \quad 1 - \beta = \frac{2}{k(k-1)} \sum_{i < j} \text{sim}(t_{ij}).$$

Step 1: Feasible range of the average intersection. By Cauchy–Schwarz,

$$\sum_x c_x^2 \geq \frac{(\sum_x c_x)^2}{U} = \frac{(kn)^2}{U} = knr,$$

hence

$$\bar{t}_{\min} = \frac{n(r-1)}{k-1}.$$

For a matching upper envelope, concentrate as many c_x ’s at k as possible and set the rest to 1. If t elements have $c_x = k$ and all others have $c_x = 1$, the constraint $\sum_x c_x = kn$ gives $t = \frac{kn - U}{k - 1} = \frac{kn(r-1)}{r(k-1)}$, and every pair shares exactly these t elements, so

$$\bar{t}_{\max} = \frac{kn(r-1)}{r(k-1)}.$$

Clearly $\bar{t}_{\min} \leq \bar{t}_{\max}$ with equality only at $r \in \{1, k\}$.

Step 2: Bounds for the average similarity $1 - \beta$. The map $\text{sim}(t) = \frac{t}{2n - t}$ is increasing and convex on $[0, n]$. By Jensen,

$$1 - \beta = \frac{2}{k(k-1)} \sum_{i < j} \text{sim}(t_{ij}) \geq \text{sim}(\bar{t}) \geq \text{sim}(\bar{t}_{\min}),$$

and the “ k -shared-core + disjoint-uniques” construction achieves the upper envelope $1 - \beta = \text{sim}(\bar{t}_{\max})$. Therefore,

$$\frac{r-1}{2(k-1) - (r-1)} \leq 1 - \beta \leq \frac{k(r-1)}{k(r+1) - 2r}, \quad r = \frac{k}{\gamma}.$$

Step 3: Bounds expressed purely via γ . Substituting $r = \frac{k}{\gamma}$ and simplifying yields closed forms:

$$\frac{k - \gamma}{2k\gamma - \gamma - k} \leq 1 - \beta \leq \frac{k - \gamma}{k + \gamma - 2}, \quad \gamma \in [1, k].$$

Equivalently, for the dissimilarity itself,

$$\frac{2(\gamma - 1)}{k + \gamma - 2} \leq \beta \leq \frac{2k(\gamma - 1)}{2k\gamma - \gamma - k}, \quad \gamma \in [1, k].$$

See Figure 5 in Appendix D for a visualization of the feasible region for different values of k .

Edge cases. At $\gamma = 1$ (all hypotheses make identical predictions on every input), both bounds give $\beta = 0$. At $\gamma = k$ (all predictions are pairwise disjoint on every input), both bounds give $\beta = 1$. Thus the bounds are tight at the extremes.

Dropping the equal-size assumption $|\mathbb{P}_{f_i}| = n$ requires replacing $\text{sim}(t)$ by the general Jaccard formula $t/(n_i + n_j - t)$ and tracking per-pair sizes (n_i, n_j) . The same extremal principles still apply; small integrality effects only perturb the finite-sample envelopes by $O(1/n)$.

G PROMPTS

Initialization prompt P_{init}

You must return one tuple of two raw strings (no Markdown fences, no back-ticks).

```
element 0 = concise natural-language hypothesis
element 1 = FULL Python source of exactly one top-level "def"
```

Code rules (apply to the source string in element 1)

- built-ins only (do not import anything)
- spaces-only indentation (4 spaces), "\n" newlines (no "\r")
- every control-flow header (if/for/while/else/elif/with/try) must break onto the next line; never place another statement after a colon
- at most 80 characters per line
- the file must compile with `ast.parse()` and execute with `exec()` unchanged
- do not add prints, tests, or extra defs; a return must appear in the function
- logic must generalize beyond the given pairs; no hard-coding

Task

- Below are (input, output) pairs \O_n .
- Infer one rule consistent with all pairs and write a function that follows it on unseen inputs.

Pairs: {{OBS_PAIRS}}

Return only:

```
(
  "My hypothesis in one sentence ...",
  "def f(x):\n    # your code\n    return y"
)
```

Example format (strictly follow):

```
(
  "Return 6 if 6 appears, else 0",
  def f(x):\n
    if 6 in x:\n
      return [6]\n
    return [0]"
)
```

Note: This template is dataset-agnostic; only \O_n is instantiated per task.

Iterative prompt P_{iter}

```

Return one tuple of two raw strings (no Markdown fences, no back-ticks).

    element 0 = concise description of a new hypothesis
    element 1 = FULL Python source of exactly one top-level "def"

Code rules (identical to  $P_{\text{init}}$ )
- built-ins only; spaces-only indentation (4); "\n" newlines
- control-flow headers must break onto the next line; no statements after colon
- <= 80 chars per line; must compile with ast.parse() and run with exec()
- no prints/tests/extra defs; the function must contain a return
- generalize beyond the pairs; no hard-coding

You have proposed the following hypotheses so far:
 $F_{\{t-1\}} = \{f_1, \dots, f_{t-1}\}$  (summaries below)
 $\{\{\text{PREVIOUS\_HYPOTHESES}\}\}$  # e.g., bullet list of brief natural-language hypotheses

Re-examine the (input, output) pairs  $\backslash O_n$  :
 $\{\{\text{OBS\_PAIRS}\}\}$ 

Your goal
- Invent a brand-new hypothesis  $f_t$  that
  (i) is consistent with all pairs in  $\backslash O_n$ , and
  (ii) is distinct in underlying principle from every  $f$  in  $F_{\{t-1\}}$ .

Return exactly:

(
  "Concise description of the new hypothesis",
  "def f(x):\n    # your code\n    return y"
)

Example format (strictly follow):

(
  "Return 6 if 6 appears, else 0",
  def f(x):\n
    if 6 in x:\n
      return [6]\n
    return [0]"
)

Note: This template is shared across datasets; only  $\backslash O_n$  and  $F_{\{t-1\}}$  vary.

```

H USE OF LARGE LANGUAGE MODELS

We used large language models solely for light language editing—polishing phrasing and detecting grammar/spelling/typos. No ideas, analyses, algorithms, datasets, annotations, or paper results were generated or modified by LLMs. All technical content was authored and verified by the authors, and no non-public or personal data were shared with LLMs.