Causal Probability-inspired Experiential Learning

Anonymous ACL submission

Abstract

Experiential learning enables LLM agents to improve from past reasoning and action trajectories, but existing methods select trajectories for insight extraction and in-context demonstration based on environmental feedback, which favors spurious associations over genuine causal ones. This paper introduces a causal probability-inspired framework that measures the probability of necessity (PN) and probability of sufficiency (PS) of individual reasoning steps. Using counterfactual interventions, we estimate the causal contribution of intermediate steps in the trajectory via a simple yet effective approximation of PN and PS, avoiding the prohibitive cost of building causal graphs or fine-tuning models. During experience collection, we first prune steps with both low PN and low PS from each trajectory, retaining only the causal backbone before extracting insights. During inference, after retrieving candidate in-context demonstrations, we re-rank them by aggregating PN and PS scores to prioritize trajectories with stronger causal contributions. Experiments on mathematical reasoning, question answering, and embodied AI tasks show that our method achieves the lowest average reasoning steps among baselines while simultaneously improving accuracy in most experiments, functioning as a plug-and-play, training-free refinement for existing experiential learning.

1 Introduction

Large language models (LLMs) (Radford et al., 2018; Guo et al., 2025a) exhibit remarkable performance on cognitively demanding tasks such as commonsense reasoning, mathematical problem-solving, and code generation. However, their competence is rooted in deep statistical fitting to co-occurrence patterns in training data, not in formal causal mechanisms. This gap between correlation and causation directly contributes to the *hallucination* problem (Sun et al., 2026): without a principled notion of causal necessity, models may omit essential logical links, leading to omission hallucinations; without causal sufficiency, models may rely on spurious correlations and fabricate unsupported content (Guo et al., 2025b).

To enable learning from past reasoning attempts, experiential learning frameworks such as ExpeL (Zhao et al., 2024) autonomously collect successful and failed trajectories, using the successful ones as in-context examples. They then abstract insights from batches of successful trajectories and extract lessons by contrasting successful and failed pairs. Crucially, ExpeL relies on environmental feedback to classify trajectories as successful or failed. This feedback captures patterns that happen to co-occur with success, rather than those that *cause* it. The core research question we address is: How can experiential learning agents measure the true causal contribution of reasoning steps? Our method answers this with a causal-probability-inspired metric that extracts backbone reasoning steps and selects in-context demonstrations.

Causality has been introduced into LLM reasoning through two main approaches. One line operates during training, using adversarial or counterfactual preference objectives (Paul et al., 2024; Rafailov et al., 2023) or causal-deconfounded fine-tuning (Wang et al., 2024). These methods, however, require task-specific labeled data and costly parameter updates, suffer from catastrophic forgetting, and transfer poorly across domains. The other line operates at inference time, constructing explicit causal graphs (Li et al., 2026; Zhang et al., 2024) or leveraging LLM-augmented causal discovery (Khatibi et al., 2024; Havrilla et al., 2025). While avoiding fine-tuning, these methods still depend on domain knowledge to build a valid causal DAG (Aziz and Brookhart, 2025) and face prohibitive computational costs as the graph scales (Yin et al., 2025).

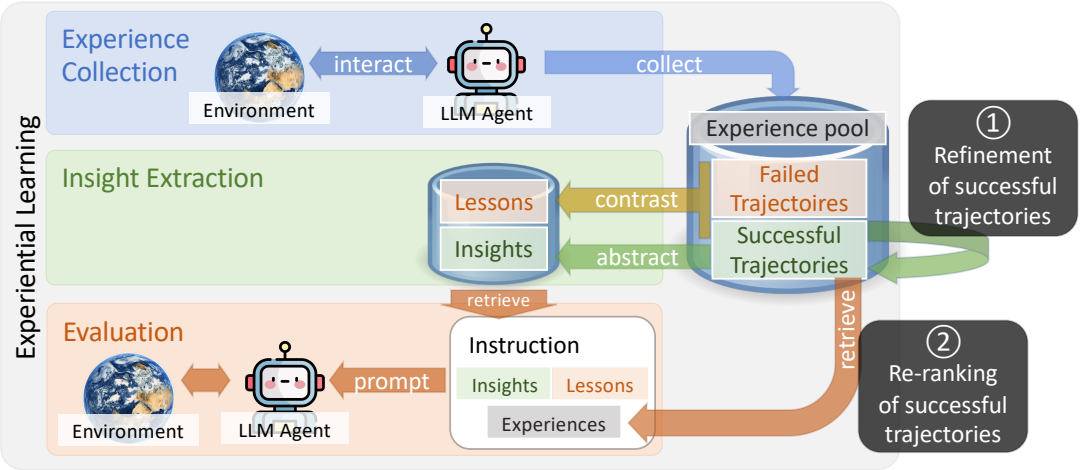

Figure 1: Illustration of a general experiential learning framework and where our trajectory refinement and re-ranking are introduced.

In this paper, we introduce a *causal-probability-inspired* (rather than strictly causal-inferential) framework that integrates the Probability of Necessity (PN) and Probability of Sufficiency (PS) (Pearl, 2000; González and Nori, 2024) into the experiential learning pipeline. We do not require a pre-specified causal graph; instead, we rely on counterfactual reasoning over natural-language reasoning steps to approximate causal contributions. Specifically, PN guides the extraction of a causally necessary backbone: steps whose removal would break the logical chain, thereby addressing omission hallucinations. PS identifies causally sufficient steps, those whose inclusion stably suffices for success. We use PN and PS to extract reasoning backbone steps and to re-rank in-context demonstrations, reducing redundancy and fabrication hallucinations. The whole framework is model-agnostic, training-free, and can be plugged into any experiential learning framework (e.g., ExpeL, Reflexion) as a drop-in experience refinement. Our main contributions are:

- We introduce a causal-probability-inspired framework that integrates PN and PS into the experiential learning loop. It replaces holistic analysis with fine-grained, counterfactual-based measurement of reasoning trajectories, enabling training-free and graph-free experience refinement and re-ranking.

- Evaluations across mathematical reasoning, commonsense QA, and embodied AI benchmarks show that our method simultaneously reduces average reasoning steps and improves accuracy over strong baselines, validating its effectiveness as a generic, plug-and-play experiential learning tool.

2 Causal Probability-Inspired Experience Refinement

We propose a causal probability-inspired experience refinement and re-ranking mechanism that integrates seamlessly into existing experiential learning frameworks such as ExpeL (Zhao et al., 2024). The key idea is to enhance environmental feedback-based selection of successful trajectories with a principled measurement of the causal contribution of individual reasoning steps. We adopt two counterfactual quantities from Pearl's causal hierarchy (Pearl, 2000): the *Probability of Necessity* (PN) and the *Probability of Sufficiency* (PS). PN identifies steps whose absence would break the logical chain, while PS quantifies how stably a step pattern suffices for success. Both quantities are estimated via counterfactual interventions over natural-language reasoning steps, requiring no explicit causal graph or model fine-tuning.

2.1 Overview of the Experience Refinement Pipeline

Experiential learning generally adopts a three-stage architecture: *experience collection*, where the agent autonomously explores training tasks and stores trajectories; *insight extraction*, where natural-language insight and lessons are distilled from past trajectories; and *evaluation*, where the agent tackles unseen tasks using the accumulated

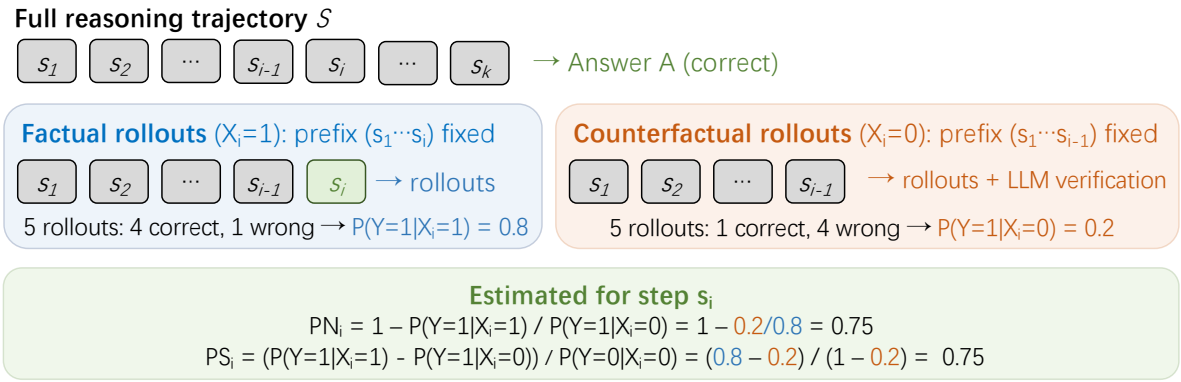

Figure 2: Example estimation of step *probability of necessity* (PN) and *probability of sufficiency* (PS).

insights and retrieved examples. As illustrated in Figure 1, our contribution lies in the introduction of causal PN and PS at the following two critical points:

Trajectory refinement for insight generation. When abstracting insights from successful trajectories, only steps with either PN or PS above predefined thresholds are retained, and insights are derived exclusively from these steps. This ensures that insights are rooted in genuinely necessary and sufficient reasoning rather than spurious co-occurrences.

Trajectory re-ranking for in-context example selection. At evaluation time, the agent retrieves candidate trajectories from the successful experience pool using semantic similarity. We re-rank these candidates using a score that combines PN and PS, promoting trajectories whose strategy patterns reliably lead to correct outcomes across varied contexts.

The overall pipeline is model-agnostic, training-free, and does not assume any pre-specified causal graph.

2.2 Counterfactual Estimation of PN and PS

We structure each reasoning trajectory as a sequence of discrete steps by introducing `<step>` tags through in-context demonstration. Consider a reasoning trajectory $S = (s_1, s_2, \ldots, s_k)$ that leads to a final answer A. Each s_i is a natural-language reasoning or action step. We treat s_i as a binary intervention variable: $X_i = 1$ if step s_i is actively produced, and $X_i = 0$ if it is suppressed. We are interested in how the presence or absence of s_i causally affects the probability of a correct outcome $Y = 1$.

Under the assumptions of monotonicity (adding a reasoning step never harms correctness) and no confounding (the step's effect is independent of unobserved factors after conditioning on the preceding context), PN and PS can be expressed in terms of observable conditional probabilities (Pearl, 2000):

$$\mathrm{PN}_i = 1 - \frac{P(Y=1 \mid X_i=0)}{P(Y=1 \mid X_i=1)},$$
$$\mathrm{PS}_i = \frac{P(Y=1 \mid X_i=1) - P(Y=1 \mid X_i=0)}{P(Y=0 \mid X_i=0)} \quad (1)$$

To estimate these probabilities, we perform counterfactual sampling while preserving the full trajectory context. For a given step i:

Factual sample ($X_i = 1$): we fix the prefix $(s_1, \ldots, s_i)$ and let the LLM freely generate the remainder of the trajectory. The empirical accuracy over multiple rollouts yields $\hat{P}(Y = 1 \mid X_i = 1)$.

Counterfactual sample ($X_i = 0$): We fix the prefix $(s_1, \ldots, s_{i-1})$ and instruct the model to continue as if s_i were absent, but we do not explicitly forbid its generation. A generation is considered a valid counterfactual rollout only if it does not contain s_i or its semantic equivalent, as determined by an LLM-based verifier. This design choice is deliberate: instructing the model to "avoid s_i" would impose an auxiliary task demand, suppressing a specific reasoning step, which itself can impair performance on the primary reasoning task; prior work shows that increased task demands systematically lower LLM performance across reasoning benchmarks, regardless of the underlying capability of interest (Hu and Frank, 2024).

Moreover, negation instructions such as "do not mention" paradoxically increase the accessibility of the forbidden concept in LLM internal representations, a phenomenon known as ironic rebound (Mann et al., 2025). Such rebound would contaminate the counterfactual rollout by activating

s_i internally while instructing its omission, artificially depressing $\hat{P}(Y = 1 \mid X_i = 0)$ and inflating PN estimates. By forbidding neither the step nor its semantic content, we avoid introducing these confounding artifacts.

Crucially, because we evaluate PN and PS by re-simulating the entire future trajectory after an intervention, we capture joint causal effects among steps. A step that is not necessary in isolation may still be part of a jointly necessary pair; our trajectory-level measurement reflects whether the accuracy drops when that step is removed, regardless of how subsequent steps compensate.

2.3 Trajectory Refinement

Given a set of successful trajectories on training tasks, we compute $\widehat{\text{PN}}_i$ and $\widehat{\text{PS}}_i$ for each step i in each trajectory. A step is considered *causally necessary* if $\widehat{\text{PN}}_i > \lambda_N$ and *causally sufficient* if $\widehat{\text{PS}}_i > \lambda_S$, where λ_N and λ_S are thresholds chosen via validation. For each successful trajectory, we extract its *causal backbone*, which consists of steps with either PN or PS above the thresholds, and discard all other steps. These backbones serve two purposes:

Insight abstraction: Only backbone steps are fed to the insight extraction module. This focuses the resulting natural-language insights on causal reasoning elements and reduces noise from filler steps.

Experience pruning: Trajectories that do not contain any backbone step are filtered out entirely; they are considered to have succeeded due to spurious correlations.

2.4 Trajectory Re-Ranking

For each retained trajectory, we keep the $\widehat{\text{PN}}_i$ and $\widehat{\text{PS}}_i$ of its backbone steps. To rank trajectories for in-context selection, we define a causal score:

$$\text{Score}(S) = \frac{1}{|\mathcal{B}(S)|} \sum_{i \in \mathcal{B}(S)} (\widehat{\text{PN}}_i + \widehat{\text{PS}}_i),$$

where $\mathcal{B}(S)$ is the set of backbone steps in S. This score rewards trajectories whose steps are both causally necessary and sufficient, i.e., they stably lead to correct answers regardless of the surrounding context.

During evaluation on a new task, we first retrieve the top-k candidate trajectories from the experience pool using Faiss-based semantic similarity (Douze et al., 2024). We then re-rank these k candidates by their causal score and select the top k' ($k' \leq k$) as in-context examples. This re-ranking prioritizes trajectories whose success is causally robust rather than merely correlational, reducing fabrication hallucinations and the influence of spurious features.

Table 1: Dataset size and max reasoning steps.

Benchmark	Size	Max Steps
HotpotQA	100	7
ALFWorld	134	20
WebShop	100	15
FEVER	100	7
MATH-500	500	7
AIME 2025/2026	60	10
ARC-Challenge	1119	7
CommonsenseQA 2.0	1434	7

2.5 Plug-and-Play Integration

Our method is designed as a drop-in module for experiential learning. It makes no assumption about the underlying LLM, requires no parameter updates, and does not rely on any pre-built causal DAG. The only domain-specific elements are the thresholds λ_N and λ_S, which can be set via light calibration on a small set of training tasks. In all experiments, we integrate the PN/PS refinement into ExpeL (Zhao et al., 2024), but the same mechanism can be applied to Reflexion (Shinn et al., 2023) or similar frameworks.

3 Related Work

3.1 LLM Reasoning and Experiential Learning

Chain-of-Thought (CoT) (Wei et al., 2022) and its extensions (Tree-of-Thoughts (Yao et al., 2023), Graph-of-Thoughts (Besta et al., 2024), self-consistency (Wang et al., 2023)) improve multi-step problem solving, yet remain grounded in statistical surface patterns rather than causal structure. Even advanced models suffer sharp accuracy drops when the wording of reasoning rules is lightly perturbed (Kiciman et al., 2023), and recent studies reveal pervasive shortcut reasoning (Cao et al., 2026) and unfaithful chains (Han et al., 2026). Experiential learning frameworks such as ExpeL (Zhao et al., 2024), ERL (Allard et al., 2026), and EvolveR (Wu et al., 2025) extend this paradigm: agents accumulate success/failure trajectories, contrast them, and distill natural-language lessons. Because success is determined by environmental feedback, the resulting selection of in-context examples and strategies

Table 2: Accuracy↑ (%) across benchmarks and LLMs. We highlight the best and second-best results.

Method	HotpotQA	ALFWorld	WebShop	FEVER	MATH-500	AIME25/26	ARC-C	CSQA2
Qwen3-14B								
CoT	21.0	20.2	7.0	30.0	45.0	6.7	73.5	76.8
ReAct	23.0	26.9	13.0	33.0	47.8	6.7	80.3	79.2
Reflexion	26.0	26.1	12.0	34.0	54.6	11.7	81.8	80.5
ExpeL	28.0	30.6	15.0	36.0	63.0	11.7	82.9	83.4
CEL	30.0	31.3	13.0	37.0	66.2	10.0	82.8	86.0
Qwen3-32B								
CoT	23.0	26.9	11.0	36.0	55.6	10.0	81.7	80.4
ReAct	27.0	32.8	16.0	40.0	58.2	11.7	84.6	84.5
Reflexion	30.0	38.1	20.0	42.0	68.2	11.7	85.9	84.3
ExpeL	35.0	35.8	18.0	44.0	72.6	13.3	87.2	89.3
CEL	34.0	38.8	18.0	46.0	71.4	13.3	90.2	93.1
DeepSeek-V3								
CoT	31.0	41.0	29.0	41.0	71.0	26.7	84.2	85.1
ReAct	34.0	46.3	32.0	47.0	73.4	28.3	85.1	88.4
Reflexion	39.0	55.2	38.0	49.0	72.8	31.7	86.4	90.9
ExpeL	46.0	60.5	36.0	55.0	80.0	30.0	91.2	92.0
CEL	51.0	65.7	42.0	61.0	84.6	36.7	94.7	96.6
GLM-4.6								
CoT	32.0	39.6	26.0	43.0	64.2	25.0	84.7	85.3
ReAct	37.0	48.5	30.0	46.0	67.0	26.7	86.5	87.5
Reflexion	43.0	56.7	32.0	51.0	69.8	28.3	87.2	88.2
ExpeL	53.0	54.5	38.0	53.0	76.4	28.3	93.3	91.2
CEL	57.0	61.2	41.0	58.0	81.2	30.0	97.0	94.0

remains correlational, often favoring spurious associations over genuine causal contributions.

3.2 Causal Methods in LLM Reasoning

Integrating causality into LLMs has followed two main approaches. Training-time interventions include counterfactual preference optimization (Paul et al., 2024; Rafailov et al., 2023) and causal-deconfounded fine-tuning (Wang et al., 2024); these demand labeled data, costly parameter updates, and suffer from catastrophic forgetting. Inference-time methods construct explicit causal graphs (Li et al., 2026; Zhang et al., 2024) or employ LLM-augmented causal discovery (Khatibi et al., 2024; Havrilla et al., 2025), but require substantial domain expertise to build a valid causal DAG, limiting scalability.

3.3 PN and PS for LLM Reasoning

Probabilities of necessity and sufficiency from Pearl's causal hierarchy (Pearl, 2000) have been recently applied to LLMs. Gonzalez et al. (González and Nori, 2024) use PN and PS as evaluation tools, showing that GPT models exhibit illusory reasoning competence.

Yu et al. (Yu et al., 2026) estimate step-level PN/PS to improve chain-of-thought, but their approach suffers from two key biases: (1) forcing the model to avoid a step during generation imposes an off-policy hard constraint, which artificially depresses $P(Y = 1|X_i = 0)$ and overestimates sufficiency; (2) their iterative pruning greedily removes steps in isolation, ignoring joint interactions where multiple steps collectively contribute to necessity. Our framework differs fundamentally: we perform rollout-then-verify counterfactual measurement, generating freely without prohibition, then filtering valid counterfactuals via an LLM verifier, which avoids off-policy distortion and captures trajectory-level joint effects. We use PN and PS not for direct step scoring but as refinement and reranking signals within an experiential learning loop. Notably, our method is plug-and-play, requires no retraining or causal graph, and directly addresses omission and fabrication hallucinations (Guo et al., 2025b).

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate our method, *Causality-inspired Experiential Learning* (CEL), on eight representative reasoning benchmarks spanning diverse domains: HotpotQA (Yang et al., 2018) for multi-hop question answering, ALFWorld (Shridhar et al., 2021) for embodied household tasks, WebShop (Yao et al., 2022a) for simulated e-commerce navigation, FEVER (Thorne et al., 2018) for fact verification, MATH-500 for mathematical problem-solving, AIME 2025 and 2026 (AIME25/26) for competition-level mathematics, ARC-Challenge

Table 3: Steps↓ and tokens↓ across benchmarks and LLMs. We highlight best and second-best results.

Method	HotpotQA		ALFWorld		WebShop		FEVER		MATH-500		AIME25/26		ARC-C		CSQA2	
	steps	tokens	steps	tokens	steps	tokens	steps	tokens	steps	tokens	steps	tokens	steps	tokens	steps	tokens
Qwen3-14B																
CoT	4.2	860.2	19.6	3008.3	14.9	1203.2	4.3	1632.2	6.7	1498.5	7.3	1403.4	3.4	764.5	3.0	751.2
ReAct	4.9	783.6	18.7	3003.6	14.2	1130.6	4.6	1689.5	7.1	1620.4	8.7	1600.6	3.4	780.0	3.0	846.1
Reflexion	4.1	660.5	18.8	2621.2	13.8	983.5	3.9	1501.4	6.6	1687.9	8.5	1555.2	4.0	876.1	3.6	702.4
ExpeL	3.9	664.1	17.6	2633.7	13.5	1040.0	4.2	1552.2	5.7	1432.0	7.5	1580.7	3.2	596.3	2.9	496.7
CEL	4.1	551.4	16.1	2270.6	14.0	920.2	3.8	1003.1	5.4	1302.0	7.5	1241.5	3.0	554.7	2.7	445.9
Qwen3-32B																
CoT	4.1	784.5	18.7	2803.6	13.5	1273.0	4.2	1365.3	6.0	1778.5	7.1	1470.4	3.2	741.5	3.0	786.0
ReAct	4.6	804.2	19.0	2852.8	14.3	1320.8	4.1	1530.9	6.9	1687.2	7.4	1458.5	3.4	800.7	3.1	832.4
Reflexion	5.1	603.8	17.6	2538.5	13.0	1048.0	4.2	1214.3	6.4	1525.8	7.3	1405.8	3.2	705.6	3.5	701.5
ExpeL	3.9	705.5	18.0	2732.4	12.4	987.0	4.2	1121.1	5.9	1497.9	7.5	1507.6	3.4	721.4	3.0	546.0
CEL	3.8	680.0	17.2	2384.4	11.8	830.5	3.4	967.5	6.0	1350.8	6.2	1200.5	3.2	647.7	2.2	485.1
DeepSeek-V3																
CoT	3.7	904.5	17.5	3324.6	14.6	1398.5	3.5	1852.2	6.4	1750.4	6.6	1686.4	2.4	1002.4	2.6	994.4
ReAct	4.1	812.5	16.2	3479.5	13.1	1530.7	3.9	1932.4	6.3	1854.2	6.8	1602.7	2.7	1204.9	2.9	1002.5
Reflexion	4.2	802.9	18.2	3497.2	13.6	1487.2	3.1	1673.5	6.1	1702.4	6.9	1680.6	3.6	910.5	3.3	660.5
ExpeL	3.6	754.6	19.5	3618.3	14.5	1207.0	3.0	1823.3	6.0	1655.0	6.3	1712.5	2.3	984.4	2.8	693.9
CEL	3.3	693.0	16.8	3057.2	11.8	1342.9	2.8	1582.4	5.2	1588.5	5.2	1565.1	2.3	892.1	2.1	705.0
GLM-4.6																
CoT	3.2	1002.5	15.4	3549.4	13.7	1693.9	3.7	1920.4	5.5	1684.2	6.1	1804.5	2.5	1245.7	2.6	1000.9
ReAct	4.5	1150.7	14.9	3939.4	13.1	1588.4	3.7	1798.4	5.7	1854.4	6.3	2004.6	2.8	1386.1	3.0	1104.1
Reflexion	4.5	984.6	15.2	3787.6	12.6	1410.8	3.1	1543.0	5.8	1610.6	6.7	1810.6	3.7	1105.8	3.4	805.5
ExpeL	3.7	900.2	15.4	3853.5	11.8	1487.3	3.3	1683.1	5.9	1621.7	6.7	1768.2	2.3	1000.6	2.6	854.4
CEL	3.9	768.2	14.0	3682.0	10.1	1289.2	3.0	1386.8	5.6	1587.6	6.3	1605.4	2.4	1008.7	2.0	790.4

(ARC-C) for grade-school science reasoning, and CommonsenseQA 2.0 (CSQA2) (Talmor et al., 2022) for commonsense reasoning. Dataset sizes and per-task maximum reasoning steps mostly follow the settings in Expel (Zhao et al., 2024) and are listed in Table 1.

Baselines. We compare against four methods spanning the spectrum from simple prompting to complex experiential learning: CoT (Wei et al., 2022), ReAct (Yao et al., 2022b), Reflexion (Shinn et al., 2023) with a single reflection round, and ExpeL (Zhao et al., 2024). ExpeL serves as our primary baseline, as our method is integrated into its experiential learning pipeline.

LLMs. We experiment with four LLMs of varying scale and capability: Qwen3-14B, Qwen3-32B (Yang et al., 2025), DeepSeek-V3 (Guo et al., 2025a), and GLM-4.6 (Hong et al., 2025). All models are accessed via Aliyun (ali) API and run with built-in thinking modes disabled to avoid confounding from implicit reasoning mechanisms.

Metrics. We evaluate both effectiveness and efficiency through three metrics. *Accuracy* uses exact match: the model output matches the ground-truth answer exactly. *Average reasoning steps* counts the number of intermediate reasoning steps produced before the final answer. *Token consumption* measures the number of output tokens generated on inferencing.

Hyperparameters. The necessity threshold is set to $\lambda_N = 0.6$ and the sufficiency threshold to $\lambda_S = 0.8$, selected based on our validation experiments. Per-task maximum reasoning steps are given in Table 1. All experiments use 5 counterfactual rollouts per step for PN/PS estimation.

4.2 Main Results

Table 2 reports the accuracy of all methods across benchmarks and LLMs, and Table 3 reports the corresponding average reasoning steps and token consumption. We highlight three key findings.

Our method achieves the best or second-best accuracy in the majority of experiments. Across all 32 comparisons (8 benchmarks $\times$ 4 LLMs), our method ranks first in 26 and second in the most of the remainder. On the stronger LLMs DeepSeek-V3 and GLM-4.6, our method leads across all 16 experiments, with average absolute improvements of $+4.74$ and $+3.68$ percentage points over the best baseline, respectively. On smaller models (Qwen3-14B, Qwen3-32B), gains are more modest and occasionally our method trails ExpeL by a narrow margin (e.g., HotpotQA on Qwen3-32B:

Table 4: Ablation accuracy↑ (%). We highlight best and second-best results.

Method	HotpotQA	ALFWorld	WebShop	FEVER	MATH-500	AIME	ARC-C	CSQA2
				Qwen3-14B				
ExpeL	28.0	30.6	15.0	36.0	63.0	11.7	82.9	83.4
CEL	30.0	31.3	13.0	37.0	66.2	10.0	82.8	86.0
– PN	29.0	31.8	13.0	36.0	64.5	11.7	82.5	82.8
– PS	31.0	31.2	14.0	37.0	65.1	10.0	83.5	86.9
				Qwen3-32B				
ExpeL	35.0	35.8	18.0	44.0	72.6	13.3	87.2	89.3
CEL	34.0	38.8	18.0	46.0	71.4	13.3	90.2	93.1
– PN	35.0	38.4	18.0	45.0	71.8	13.3	88.6	91.5
– PS	34.0	36.0	19.0	44.0	72.0	13.3	88.0	92.8
				DeepSeek-V3				
ExpeL	46.0	60.5	36.0	55.0	80.0	30.0	91.2	92.0
CEL	51.0	65.7	42.0	61.0	84.6	36.7	94.7	96.6
– PN	47.0	64.4	37.0	58.0	82.5	31.7	93.2	94.8
– PS	50.0	62.7	40.0	59.0	83.1	33.3	94.0	94.0
				GLM-4.6				
ExpeL	53.0	54.5	38.0	53.0	76.4	28.3	93.3	91.2
CEL	57.0	61.2	41.0	58.0	81.2	30.0	96.9	93.9
– PN	56.0	60.2	41.0	55.0	79.2	28.3	96.3	92.9
– PS	57.0	59.5	39.0	56.0	78.5	30.0	94.8	92.5

34.00% vs. 35.00%). Our manual analysis reveals that smaller models frequently produce malformed outputs that fail to parse, inflating the failure count.

Our method reduces reasoning steps and reduces token consumption. In 22 of 32 experiments, our method achieves the lowest average number of reasoning steps; in the remaining cases it closely matches the best baseline. With DeepSeek-V3, for example, our method reduces average steps by 1.1 steps (17.5%) over ExpeL on AIME25/26, by 0.67 steps (23.8%) on CommonsenseQA 2.0, and by 11% on WebShop. With smaller models such as Qwen3-14B, step reductions are observed on some tasks (e.g., MATH-500: -0.25 steps). Token consumption follows a similar pattern: our method achieves the lowest token count in 27 of 32 experiments. These results validate that our causal-inspired mechanism effectively prunes causally redundant reasoning steps.

Our method generalizes across diverse LLMs and task domains. With DeepSeek-V3, the largest improvement of our CEL over ExpeL is on FEVER (6%), followed by ALFWorld 5.22%, AIME25/26 6.7%, and MATH-500 4.6%. On smaller models, the advantage narrows but remains positive overall: with Qwen3-32B, CEL outperforms ExpeL in 6 of 8 tasks. Tasks involving rigorous logical reasoning (FEVER, AIME, MATH-500) may benefit substantially from PN's ability to identify causal bottlenecks; tasks with broad reasoning spaces (CommonsenseQA 2.0, ALFWorld) may benefit from PS's ability to surface robust strategies.

4.3 Ablation Study

To isolate the contributions of PN and PS, we compare the full method against two variants: *CEL without PN*, which uses only the PS values, and *CEL without PS*, which uses only the PN values. Results are reported in Table 4.

Both PN and PS contribute positively, with stronger synergy on larger LLMs. Removing either PN or PS degrades performance on most tasks, confirming that PN and PS serve complementary functions. On DeepSeek-V3 and GLM-4.6, both single-component variants (CEL without PN and CEL without PS) already outperform ExpeL, with average gains of 2.72% and 2.55% respectively, and the full method achieves the best results. On Qwen3-14B and Qwen3-32B, single-component gains are smaller and more variable, with occasional regressions (e.g., HotpotQA, WebShop). This supports our earlier observation: causal probability measurement requires sufficient model capability for reliable counterfactual simulation.

4.4 Threshold Sensitivity Analysis

We further conduct a study on the selection of λ_N and λ_S (Figure 3). The best performance concentrates around $\lambda_N \in [0.6, 0.7]$ and $\lambda_S \in [0.7, 0.8]$, reaching 0.52 accuracy. Low thresholds (≤ 0.4) retain many redundant steps (since satisfying either PN$> \lambda_N$ or PS$> \lambda_S$ suffices), dropping accuracy to 0.41–0.43. High thresholds (≥ 0.9) prune overly aggressively, reducing accuracy to 0.34. Notably, setting one threshold high and the other low (e.g., $\lambda_N = 0.8, \lambda_S = 0.2$) fails to achieve optimal

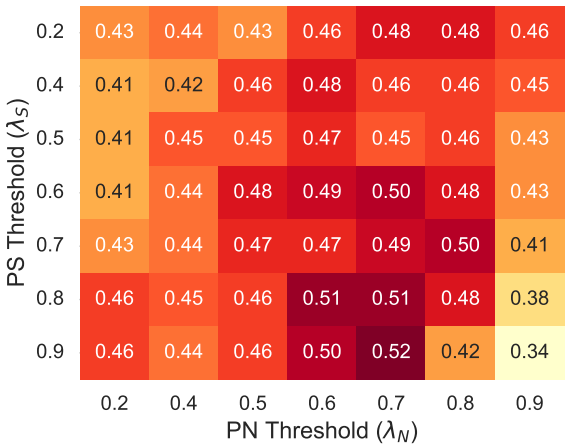

Figure 3: Accuracy under different λ_N and λ_S thresholds on HotpotQA with DeepSeek-V3.

performance, showing that while our rule requires only PN or PS to exceed its threshold, the best results arise when both thresholds are moderately high—retaining steps that are both necessary and sufficient.

4.5 Case Study

We illustrate the PN/PS backbone extraction process on a trajectory from ALFWorld for the task "heat some bread and put it in countertop." The successful trajectory consists of 9 actions (A1–A9). Table 5 shows the estimated PN and PS for each action. Using thresholds $\lambda_N = 0.6$ and $\lambda_S = 0.85$ (with $\geq$), we retain steps satisfying PN $\geq \lambda_N$ or PS $\geq \lambda_S$. This retains A1, A2, A3, A4, A5, A6, A8, A9 (8 steps). The pruned action is A7, which have low PN and PS below thresholds. These are deemed causally redundant because the heating action implicitly handles microwave operation, and closing the microwave is unnecessary. The resulting backbone preserves all causally critical nodes while reducing the trajectory length by 22.2%.

This case demonstrates the complementary roles of PN and PS: low PN identifies steps whose removal does not break the logical chain (A5, A7), while high PS marks steps that are individually sufficient for sub-goal completion. The extracted backbone is both necessary (no critical step omitted) or sufficient (each step reliably advances the task), consistent with our framework.

5 Conclusion

We introduced a causal probability-inspired framework that integrates Probability of Necessity (PN) and Probability of Sufficiency (PS) into the experiential learning pipeline of LLM agents. By replacing holistic contrastive analysis with fine-grained counterfactual measurement of reasoning steps, our method extracts causal backbones for insight generation and re-ranks retrieved trajectories by their causal scores for in-context example selection. The framework is simple to implement, requires no model fine-tuning and no pre-specified causal graph, making it a drop-in refinement for existing experiential learning agents such as ExpeL.

Experiments across eight benchmarks spanning mathematical reasoning, fact verification, question answering, and embodied AI tasks demonstrate that our method simultaneously improves accuracy and reduces reasoning steps in the large majority of experiments, with the strongest gains observed on capable LLMs where counterfactual estimation is reliable. Ablation studies confirm that PN and PS contribute complementary benefits.

Our work demonstrates that principled causalilty-inspired measurement, even approximated through counterfactual prompting, offers a practical path toward more robust and efficient experiential learning in LLM agents. We hope this exploration encourages further bridging of the gap between statistical association and causal understanding in AI reasoning.

Table 5: Case Study: PN and PS results.

Scenario: You are in the middle of a room. Looking quickly around you, you see a cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 1, a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.
Task: heat some bread and put it in countertop.

ID	Action	PN	PS	$\geq \lambda_N$	$\geq \lambda_S$
A1	go to fridge 1	0.95	0.93	✓	✓
A2	open fridge 1	0.87	0.88	✓	✓
A3	take bread 1 from fridge 1	1.00	0.93	✓	✓
A4	go to microwave 1	0.86	0.86	✓	✓
A5	open microwave 1	0.15	0.82	✗	✓
A6	heat bread 1 with microwave 1	0.95	0.98	✓	✓
A7	close microwave 1	0.42	0.68	✗	✗
A8	go to countertop 1	0.92	0.85	✓	✓
A9	put bread 1 on countertop 1	0.95	0.90	✓	✓

Limitations

We identify several limitations of the current work. First, the effectiveness of PN/PS estimation depends on the LLM's capability; as shown in our experiments, smaller LLMs yield noisy causal es-

timates, which can diminish the benefit of our method. Second, the counterfactual sampling procedure introduces computational overhead proportional to the number of rollouts per step, although this cost is incurred only during the offline experience refinement phase. Third, the thresholds λ_N and λ_S currently require light per-domain calibration; developing automatic or adaptive threshold selection mechanisms is left to future work. Finally, our evaluation is limited to relatively structured reasoning tasks with well-defined success criteria; extending causal probability measurement to open-ended generation remains an open challenge.

References

https://www.aliyun.com/.

Marc-Antoine Allard, Arnaud Teinturier, Victor Xing, and Gautier Viaud. 2026. Experiential reflective learning for self-improving llm agents. *arXiv preprint arXiv:2603.24639*.

Maryam Aziz and M Alan Brookhart. 2025. Can contemporary large language models provide the domain knowledge needed for causal inference? evaluating automated causal graph discovery through an ascvd case study. *Clinical Epidemiology*, pages 863–873.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, and 1 others. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:17682–17690.

Hongyu Cao, Kunpeng Liu, Dongjie Wang, and Yanjie Fu. 2026. Mitigating shortcut reasoning in language models: A gradient-aware training approach. *arXiv preprint arXiv:2603.20899*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Javier González and Aditya V Nori. 2024. Does reasoning emerge? examining the probabilities of causation in large language models. *Advances in Neural Information Processing Systems*, 37:117737–117761.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Peizheng Guo, Jingyao Wang, Wenwen Qiang, Huijie Guo, Changwen Zheng, Jiahuan Zhou, and Gang Hua. 2025b. Hacking hallucinations of mllms with causal sufficiency and necessity. *arXiv e-prints arXiv:2508.04182*.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. 2026. Rfeval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. *arXiv preprint arXiv:2602.17053*.

Alex Havrilla, David Alvarez-Melis, and Nicolo Fusi. 2025. Igda: Interactive graph discovery through large language model agents. *arXiv preprint arXiv:2502.17189*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. 2024. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Yuangang Li, Yiqing Shen, Yi Nian, Jiechao Gao, Ziyi Wang, Chenxiao Yu, Shawn Li, Jie Wang, Xiyang Hu, and Yue Zhao. 2026. Mitigating hallucinations in large language models via causal reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Logan Mann, Nayan Saxena, Sarah Tandon, Chenhao Sun, Savar Toteja, and Kevin Zhu. 2025. Don't think of the white bear: Ironic negation in transformer models under cognitive load. *arXiv preprint arXiv:2511.12381*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Judea Pearl. 2000. Causality: models, reasoning, and inference. *Cambridge university press*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. *International Conference on Learning Representations (ICLR)*.

Yiyou Sun, Yu Gai, Lijie Chen, Abhilasha Ravichander, Yejin Choi, Nouha Dziri, and Dawn Song. 2026. Why and how llms hallucinate: Connecting the dots with subsequence associations. *Advances in Neural Information Processing Systems*, 38:35159–35203.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819. Association for Computational Linguistics.

Ruoyu Wang, Xiaoxuan Li, and Lina Yao. 2024. Deconfounded causality-aware parameter-efficient fine-tuning for problem-solving improvement of llms. In *International Conference on Web Information Systems Engineering*, pages 161–176. Springer.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and 1 others. 2025. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Naiyu Yin, Tian Gao, and Yue Yu. 2025. Learning causal graphs at scale: A foundation model approach. *arXiv preprint arXiv:2506.18285*.

Xiangning Yu, Zhuohan Wang, Linyi Yang, Haoxuan Li, Anjie Liu, Xiao Xue, Jun Wang, and Mengyue Yang. 2026. Causal sufficiency and necessity improves chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 38:126109–126141.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19533–19541.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.