BEYOND SIMILARITY: PARTIAL INFORMATION DECOMPOSITION FOR COMPOSED SKILLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models increasingly rely on retrieving and injecting reusable skills from external libraries. The prevailing paradigm often assumes a positive correlation between skill quantity and performance gain, directly concatenating the top-ranked similar skills into the prompt context. However, our study reveals significant limitations in this approach: aggregating highly related skills frequently leads to severe diminishing returns and information redundancy, or even induces negative interference that degrades model performance. The primary contribution of this paper is the introduction of a novel analytical framework based on Partial Information Decomposition (PID) to systematically investigate the discrepancy between "retrieval similarity" and "actual compositional utility." By decomposing the model's log-likelihood into shared evidence, skill-specific evidence, and non-additive interactions, this framework enables precise quantitative evaluation of the complementary or interfering effects among multiple skills. Our extensive evaluations across seven math, question-answering, and code generation benchmarks reveal that top-ranked skill combinations typically suffer from information redundancy, yielding a marginal gain that is frequently negative relative to the best single skill. Furthermore, as an extension of this analytical framework, we conduct a preliminary exploration of using PID as a guiding strategy for skill selection; empirical results indicate that filtering skills via PID can mitigate inter-skill interference to some extent, offering a promising pathway for optimizing multi-skill composition in the future.

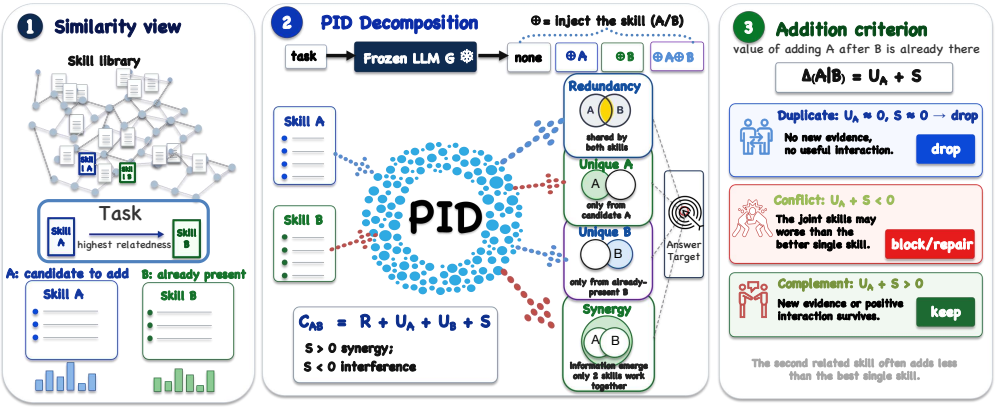


Figure 1: Overview. **(1)** Similarity retrieves candidates but not their marginal value. **(2)** No-skill, single-skill, and joint-skill probes yield redundant, source-specific, and non-additive evidence. **(3)** The atoms characterise overlap and interaction, while $\Delta_{a|b} = U_a + S$ reports the evidence gained or lost by adding a skill.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large language model systems retrieve reusable *skills*, including procedures, instructions, and workflow memories (Wang et al., 2024; Zhao et al., 2024; Wang et al., 2025c; Li et al., 2026). Many systems inject several top-ranked skills into one prompt. Related skills may repeat a procedure, reinforce an error, or become useful only in combination (Liu et al., 2026; Bai et al., 2026).

Retrieval usually ranks each skill by its similarity to the task. That score helps find plausible candidates, but it does not measure what a skill adds to an existing prompt. Two high-ranked skills may repeat the same instruction. They may also prescribe incompatible steps, or supply different pieces of a solution that work well together. These possibilities matter whenever a system injects more than one skill: the value of the second depends on what the first already provides and on how the generator uses their combination.

We therefore ask whether a second skill improves on the better standalone skill. A comparison of single-skill accuracy alone cannot answer this question. It misses both the pair's shared contribution and effects that appear only when the skills are combined. A joint accuracy score records the outcome, but does not explain whether the pair helped through complementary evidence or hurt through a change in how the model used its context. The analysis needs to separate these contributions while retaining the sign of evidence for the reference answer.

We bring partial information decomposition (PID; Williams & Beer, 2010) to prompt-level skill analysis. Treating two skills as information sources, we measure the generator's reference-answer evidence with no skill, with each skill alone, and with both skills (Figure 1). A signed decomposition separates shared, source-specific, and non-additive evidence. If skill a is added after the better standalone skill b, its evidence marginal is $U_a + S$. The identity gives a direct test of whether the addition helps, while the atoms show how the pair's contributions are distributed. For multiple-choice answers, we also use a strict Williams–Beer decomposition of predicted options.

We test this account with skills retrieved without access to the reference answer across seven math, question-answering, and code benchmarks. We then compare evidence with free-generation outcomes, including an interactive agent setting. Constructed hints with complementary content provide a positive control. Order swaps, matched-length placebo prompts, and prompt controls probe how much of the measured interaction depends on prompt construction. Together, these tests distinguish a retriever's ability to find related skills from the conditional value of placing those skills in one prompt.

Our contributions are:

- **PID for skill analysis.** We apply the PID idea to analyse prompt-level skill composition. A signed decomposition separates shared, source-specific, and interaction evidence, and expresses the added skill's marginal as $U_a + S$.
- **Broad empirical audit.** Under gold-free retrieval, the median second-skill marginal is negative on all seven benchmarks.
- **Behavioural validation and positive control.** Free-generation experiments expose dilution, while complementary hints yield positive evidence marginals.

2 RELATED WORK

Partial information decomposition. Williams & Beer (2010) decompose two sources' information about a target into redundant, unique, and synergistic atoms; later work refines these terms (Harder et al., 2013; Bertschinger et al., 2014; Griffith & Koch, 2014). Signed pointwise variants distinguish shared support from shared negative evidence (Ince, 2017; Finn & Lizier, 2018). We apply the decomposition to documents injected into prompts, building on analyses of information across modalities (Liang et al., 2023a;b).

Skill libraries and context utility. Agents accumulate reusable behaviours, workflows, and tools (Wang et al., 2024; Zhu et al., 2023; Zhao et al., 2024; Wang et al., 2025c; Yuan et al., 2024); recent work studies how to evolve, route, and benchmark these resources (Mi et al., 2026; Wang et al., 2025a; Shi et al., 2026; Wang et al., 2026a; Li et al., 2026; Zheng et al., 2026). Similarity retrieval may be followed by reranking or structured composition (Zhao et al., 2026; Liu et al., 2026; Zeng

et al., 2026). Our analysis measures the marginal value of adding one retrieved skill to another. Other work scores individual contexts (Shi et al., 2024; Liu & Wang, 2023; Yang et al., 2024; Wang et al., 2025b; 2026b) or learns when to retrieve (Jiang et al., 2023; Jeong et al., 2024; Asai et al., 2024). Competing contexts can degrade multi-document reasoning (Shi et al., 2023; Liu et al., 2024).

3 A Partial Information Decomposition for Skill Composition

3.1 Skills as Information Sources

Let G be a frozen generator, x_t a task prompt, and y_t^* its reference answer. Injecting skill s defines a conditioned predictive distribution $P_s(\cdot) = P_G(\cdot \mid x_t \oplus s)$, where $\oplus$ denotes composition. For skills X_1, X_2 and target Y, PID writes

$$I(Y; X_1, X_2) \;=\; I_{\text{red}} + I_{\text{uniq}}^{(1)} + I_{\text{uniq}}^{(2)} + I_{\text{syn}}, \tag{1}$$

where the atoms represent shared, source-specific, and jointly available information. Signed extensions also expose negative evidence.

We use signed task-level evidence for free-form answers and strict Williams-Beer PID for small discrete answer spaces.

3.2 A Signed Pointwise PID for Skill Interaction

For task t, let C_i and C_{ij} be the single- and joint-skill evidence contributions. To decompose $C_{ij} = R + U_i + U_j + S$, we define sign-consistent pointwise overlap

$$R(t) = \begin{cases} \operatorname{sgn}(C_i)\min(|C_i|, |C_j|)\,, & C_i C_j > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

and define the remaining terms residually:

$$U_i = C_i - R, \quad U_j = C_j - R, \quad S = C_{ij} - C_i - C_j + R. \tag{3}$$

The four atoms capture shared support ($R > 0$), shared negative evidence ($R < 0$), source-specific evidence (U_i), and non-additivity (S) (Ince, 2017; Finn & Lizier, 2018).

Interpretation. Identical positive contributions enter R, leaving zero unique terms. Shared decreases in target likelihood yield $R < 0$. The interaction S is positive for super-additive composition and negative for suppressive composition. A positive S can coexist with a negative marginal (Eq. 5).

Gold-evidence instance. We use teacher-forced log-likelihood gain on y_t^* (Liu & Wang, 2023; Wang et al., 2025b):

$$\begin{aligned} C_i(t) &= \log P_G(y_t^* \mid x_t \oplus s_i) - \log P_G(y_t^* \mid x_t), \\ C_{ij}(t) &= \log P_G(y_t^* \mid x_t \oplus s_i \oplus s_j) - \log P_G(y_t^* \mid x_t). \end{aligned} \tag{4}$$

For this evidence instance, we write $D_i = C_i$ and $D_{ij} = C_{ij}$ below.

Proposition 1 (Properties of the signed decomposition) *For every task t, Eqs. 2 and 3 satisfy:* (i) ***exactness***, *$C_{ij} = R + U_i + U_j + S$ by construction;* (ii) ***symmetry***, *$R(s_i, s_j) = R(s_j, s_i)$;* (iii) ***self-redundancy***, *if $C_i = C_j$ then $R = C_i$ and $U_i = U_j = 0$;* (iv) ***duplication invariance***, *injecting a copy of s_i yields no unique information; and* (v) ***sign consistency***, *R inherits the sign of the shared contribution, so $R < 0$ marks a shared decrease in reference-answer likelihood.*

These task-level atoms measure signed evidence. The weaker contribution enters R when both contributions have the same sign. Different texts can therefore produce zero unique evidence. Alternative redundancy definitions change the atom allocation but preserve the addition criterion (Appendices C and I).

3.3 Marginal Value of an Added Skill

Let s_b denote the better standalone skill and s_a the candidate addition. Its evidence marginal is

$$\Delta_{a|b} = D_{ij} - \max(D_i, D_j) = U_a + S. \tag{5}$$

Because R cancels, the marginal is invariant to the redundancy definition. Its sign records whether the addition loses or gains reference-answer evidence; the redundancy convention affects the atom allocation.

3.4 Strict Williams-Beer Decomposition on Discrete Answers

When the answer space is a small discrete set (multiple-choice tasks), the classical decomposition applies. Let X_i be the skill-conditioned predicted option and Y the gold option. From the empirical $p(x_1, x_2, y)$ we use the $I_{\min}$ redundancy of Williams & Beer (2010),

$$R_{\mathrm{wb}} = \sum_y p(y) \min_{i \in \{1,2\}} I(X_i; Y{=}y)\,, \tag{6}$$

with $U_i = I(Y; X_i) - R_{\mathrm{wb}}$ and $S_{\mathrm{wb}} = I(Y; X_1, X_2) - I(Y; X_1) - I(Y; X_2) + R_{\mathrm{wb}}$. We subtract each atom's mean under 500 label permutations to correct plug-in bias; Appendix H gives further diagnostics.

3.5 Validation Strategy

Five constructions with known structure test the estimator: REDUNDANT, UNIQUE, SYNERGY, CONFLICT, and NOISE (Appendix D). We then compare teacher-forced evidence with strict PID, free generation, injection-order swaps, and exact-length placebo controls.

4 Experimental Setup

Generator, targets, and libraries. We use frozen Qwen3-8B (Yang et al., 2025), served by vLLM (Kwon et al., 2023). The gold-evidence audit covers GSM8K (Cobbe et al., 2021), Math500 (Hendrycks et al., 2021b; Lightman et al., 2023), GPQA (Rein et al., 2024), MMLU (Hendrycks et al., 2021a), HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and BigCodeBench-Hard (Zhuo et al., 2025); MMLU enables strict PID. We distill each library from training solutions (Appendix B). Generation audits also cover HotpotQA (Yang et al., 2018), SearchQA (Dunn et al., 2017), and an interactive AlfWorld agent (Shridhar et al., 2021).

Pair construction and statistics. Leave-one-task-out TF-IDF retrieval selects three in-domain skills without using the reference answer. Their three pairs yield 300 records per dataset and 192 for HumanEval. Appendix B gives the full protocol; Appendix A.1 reports a reference-ranked stress test. Behavioural validation evaluates the no-skill, two single-skill, and joint conditions. It reports the composition marginal $\Delta^{\mathrm{beh}}_{a|b} = \mathrm{pass}(ab) - \max\{\mathrm{pass}(a), \mathrm{pass}(b)\}$. The better-single comparator is defined from the measured outcomes of each retrieved pair.

5 Diagnosing the Marginal Value of an Added Skill

5.1 A Related Second Skill Usually Fails to Add Value

We compare each joint prompt with its better single-skill prompt, then decompose the evidence difference.

Table 1 and Figure 2 show a negative median second-skill marginal on every dataset, with task-bootstrap intervals below zero. The joint prompt beats the better single in 17% to 39% of pairs. Math and science pairs show small or negative interaction. Code pairs combine negative shared evidence with positive interaction, but the joint prompt still trails the better single. Topical relatedness does not imply added evidence. A high-evidence stress test appears in Appendix A.1.

Table 1: Gold-free TF-IDF top-3 in-domain audit (medians, nats, Qwen3-8B). $\Delta_{a|b} = D_{ij} - \max(D_i, D_j)$; *joint>best* is the fraction with $\Delta_{a|b} > 0$. Each task contributes three pairs, so n counts pairs from 100 independent tasks (64 for HumanEval). Intervals resample tasks with all their pairs retained (10,000 bootstrap draws).

dataset	n	R	U_1	U_2	S	$\Delta_{a\|b}$	95% CI	joint>best
GSM8K	300	+0.00	+0.00	+0.00	−0.04	−0.36	[−0.46, −0.28]	0.25
Math500	300	−0.04	+0.00	+0.00	−0.17	−0.53	[−0.61, −0.47]	0.17
GPQA	300	+0.06	+0.00	+0.00	−0.12	−0.24	[−0.31, −0.17]	0.27
MMLU	300	−0.73	+0.00	+0.00	+0.23	−0.52	[−0.71, −0.40]	0.29
HumanEval	192	−5.30	+0.00	−0.30	+2.09	−0.65	[−1.05, −0.21]	0.36
MBPP	300	−3.58	+0.00	+0.00	+1.02	−2.96	[−3.51, −2.27]	0.21
BCB-Hard	300	−4.87	−0.18	+0.00	+3.33	−1.00	[−1.62, −0.48]	**0.39**

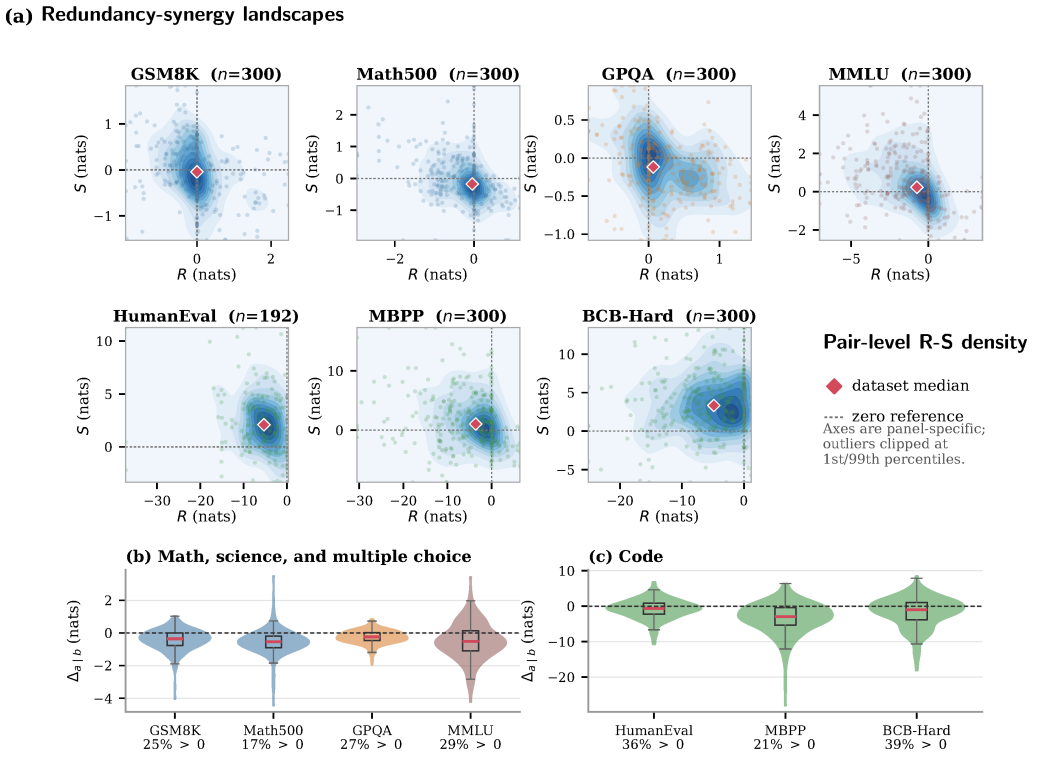


Figure 2: Gold-free TF-IDF audit across seven datasets. **Left:** density of redundancy R versus synergy S; contours enclose the highest-density regions and diamonds mark medians. **Right:** distribution of the second-skill marginal $\Delta_{a|b} = D_{ij} - \max(D_i, D_j)$.

Our overlap convention gives zero median unique atoms in most settings. Singleton effects point in the same direction for 73% to 95% of pairs. Within those pairs, the weaker effect retains 51% to 71% of the stronger magnitude (Table 20).

5.2 The Decomposition Recovers Known Structure

On tasks that the generator answers incorrectly without help, strict PID recovers the intended atom for duplicated, unique, and jointly decodable hints. Noise remains near zero (Figure 3). Table 7 includes the retrieved-library control and full values.

Reading redundancy. Although R cancels in $\Delta_{a|b}$, its sign distinguishes shared support from shared negative evidence. Separate strict checks assign 62% of corrected MMLU answer information and 77% of corrected HumanEval joint-success information to redundancy (Tables 3 and 10).

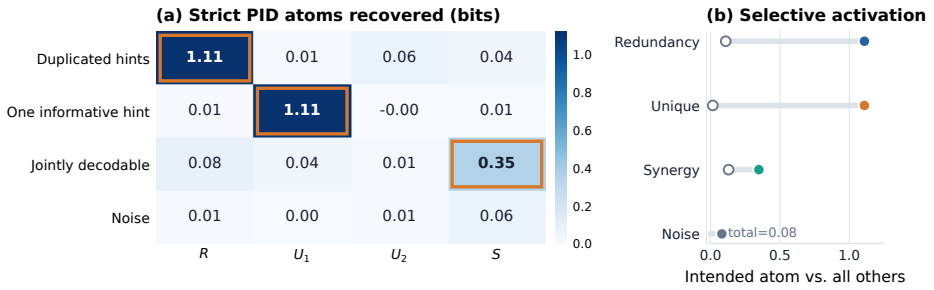


Figure 3: Oracle validation of the strict PID estimator. Each designed construction activates its intended atom: redundancy for duplicated hints, unique information for a single informative source, synergy for a jointly decodable pair, and near-zero information for noise.

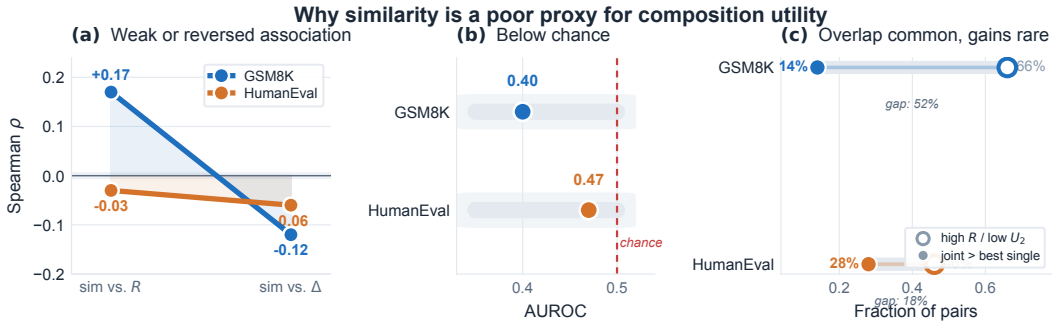


Figure 4: Similarity versus composition value. **(a)** Spearman correlations with redundancy and the second-skill marginal. **(b)** AUROC for a positive marginal, with chance at 0.50. **(c)** Fractions of pairs with high redundancy and low second-skill uniqueness, and of pairs where the joint prompt beats the better single.

5.3 Similarity Does Not Predict Marginal Value

Similarity scores do not reliably track the second-skill marginal. On IG-ranked GSM8K and HumanEval, the correlations are -0.12 and -0.06. The corresponding AUROCs for a positive marginal are 0.40 and 0.47 (Figure 4; full results in Appendix J).

5.4 Behavioural Consequences of Adding a Skill

Across ten free-generation settings, the mean composition marginal is negative in eight and never exceeds $+0.006$. The joint prompt beats the better single in at most 20% of pairs (Table 2). In repeated AlfWorld rollouts, success rises from 0.125 without a skill to 0.440 with one skill, then falls to 0.379 with two (Figure 6). In fixed-depth retrieval, GSM8K pass@1 is 0.922 at $k = 0$ and 0.909 at $k = 2$ (Figure 5); the official SkillX τ^2-Bench pipeline peaks at $k = 1$ (Figure 10).

5.5 Stability Checks

We check prompt order and second-slot context effects.

Swapping skill order preserves the negative median on all seven datasets. Exact-length controls also reveal second-slot context effects (Appendix F.2). Additional checks appear in Appendices A, E, F, and I.

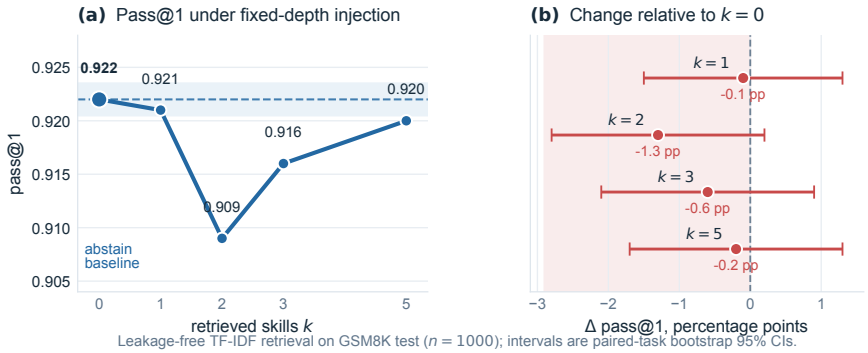


Figure 5: Leakage-free fixed-depth TF-IDF retrieval on GSM8K (n=1000, Qwen3-8B). No fixed depth improves significantly over injecting no skill; points and intervals show paired pass@1 changes from $k = 0$.

Table 2: Four-condition generation audits (Qwen3-8B). *Best* and *joint* are mean pass rates; *positive* is the fraction of pairs with a positive behavioural composition marginal. Sampling protocols vary by dataset, so comparisons are within rows.

setting	protocol	n	best	joint	$\bar{\Delta}^{\text{beh}}$ / positive
GSM8K	30 samples	100	.742	.712	−.029 / .20
MMLU	30 samples	80	.725	.677	−.048 / .06
GPQA	30 samples	100	.303	.275	−.028 / .15
HotpotQA	30 samples	100	.523	.529	+.006 / .05
SearchQA	30 samples	100	.427	.395	−.032 / .00
Math500	3 runs	30	.833	.689	−.144 / .00
MBPP	3 runs	30	.589	.500	−.089 / .00
HumanEval	3 runs	90	.615	.533	−.081 / .07
BCB-Hard	3 runs	30	.100	.100	±.000 / .07
AlfWorld	10 rollouts	100	.440	.379	−.061 / .06

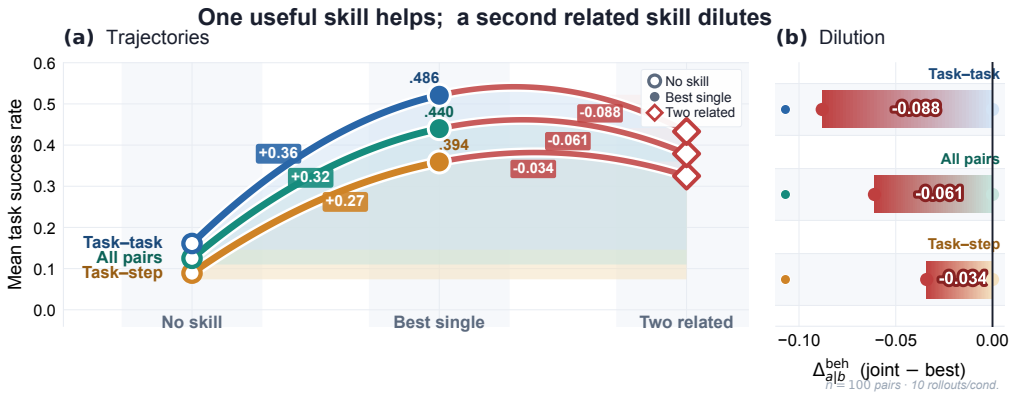


Figure 6: Interactive AlfWorld validation. **(a)** Success paths from no skill to the best single and then two related skills. **(b)** Composition marginals for two pair types and their aggregate.

5.6 A Constructed Complement Gives Positive Evidence Marginals

In a GSM8K construction, separate quantity and formula hints yield a median S of $+9.82$ nats and a positive evidence marginal on 84% of tasks. These hints supply different parts of a solution, unlike the related instructions in the primary audit.

Table 3: Strict Williams-Beer PID on MMLU (bits), aggregated over pair types. The null is the mean of 500 label permutations; corrected values subtract this mean.

atom	observed	null	corrected	p
$I(Y; X_1, X_2)$	0.284	0.014	0.270	0.002
R	0.170	0.002	**0.168**	0.002
U_1	0.000	0.001	−0.001	0.932
U_2	0.050	0.001	0.049	0.002
S	0.064	0.010	0.054	0.002

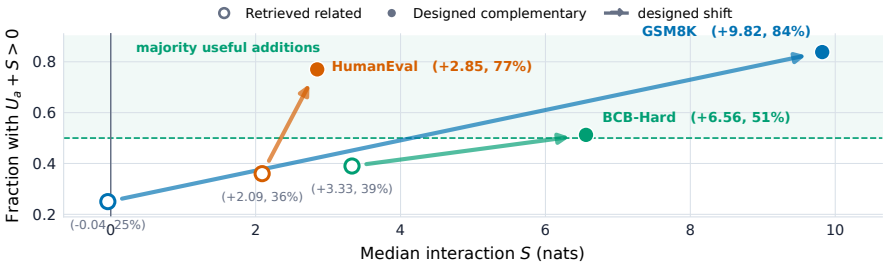


Figure 7: Related retrieval versus designed complementarity. Open circles show retrieved pairs; filled circles show constructed complementary hints. The construction moves all three domains toward larger interaction and more frequent positive second-skill marginals, separating retrieval failure from decomposition failure.

5.7 Implications for Skill Selection

Similarity can propose candidates, while the conditional marginal measures their added value. Deployment calls for a reference-free estimate of this marginal followed by generation tests. The broader lesson is that retrieval similarity is useful for proposing candidates, but conditional value remains the criterion for composition. In a focused composition check, using the PID interaction to merge related skills improved the decoded success rate from 66% to 73% on MMLU and from 48.4% to 51.6% on HumanEval. We present these as illustrative evidence that interaction-aware composition can help when the two skills are complementary, rather than as a standalone routing benchmark. Higher-order composition checks appear in Appendix G.

6 Discussion and Conclusion

Across seven benchmarks, PID separates the added skill's specific evidence from its interaction with the first skill. Retrieved pairs usually lose evidence relative to the better single, whereas complementary hints add it. Order swaps, placebo prompts, and free generation show that interaction atoms also capture generator and prompt-context effects. Skill composition should be selected by conditional value and verified on decoded outcomes, not inferred from similarity alone.

Scope and extensions. The current study covers evaluated skill pairs. Naturally authored libraries, additional generators, and longer skill chains offer direct extensions.

Ethics Statement

This work uses public benchmarks and open-weight models. It involves no human subjects or personal data and introduces no application beyond those of the underlying LLMs.

Reproducibility Statement

Experiments use frozen model weights. Section 4 and Appendix B describe skill construction, task and pair selection, inference settings, and statistical procedures. Sample counts are reported in Tables 1 and 2; Appendix C contains the proofs. Code and result artifacts will be released.

AI USE STATEMENT

We used generative AI to identify relevant references, process data, distill skills from training solutions, audit skill relations, and assist with writing. The authors verified all AI-assisted work and take responsibility for the final paper.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Tong Bai, Zhenglin Wan, Pengfei Zhou, Xingrui Yu, Yang You, and Ivor W. Tsang. SkillDAG: Self-evolving typed skill graphs for LLM skill selection at scale, 2026. URL `https://arxiv.org/abs/2606.03056`.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Conor Finn and Joseph T. Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, 20(4):297, 2018. doi: 10.3390/e20040297.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pp. 159–190. Springer, 2014. doi: 10.1007/978-3-642-53734-9_6.

Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Physical Review E*, 87(1):012130, 2013. doi: 10.1103/PhysRevE.87.012130.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021b.

Robin A. A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017. doi: 10.3390/e19070318.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks, 2026. URL https://arxiv.org/abs/2602.12670.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Dawei Liu, Zongxia Li, Hongyang Du, Xiyang Wu, Shihang Gui, Yongbei Kuang, and Lichao Sun. Graph-of-skills: Dependency-aware structural retrieval for massive agent skills, 2026. URL https://arxiv.org/abs/2604.05333.

Hongfu Liu and Ye Wang. Towards informative few-shot prompt with maximum information gain for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15825–15838, 2023. doi: 10.18653/v1/2023.findings-emnlp.1060.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Qirui Mi, Zhijian Ma, Mengyue Yang, Haoxuan Li, Yisen Wang, Haifeng Zhang, and Jun Wang. Skill-Pro: Learning reusable skills from experience via non-parametric PPO for LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.01869.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024. doi: 10.18653/v1/2024.naacl-long.463.

Yaorui Shi, Yuxin Chen, Zhengxi Lu, Yuchun Miao, Shugui Liu, Qi Gu, Xunliang Cai, Xiang Wang, and An Zhang. Skill1: Unified evolution of skill-augmented agents via reinforcement learning. *arXiv preprint arXiv:2605.06130*, 2026. doi: 10.48550/arXiv.2605.06130.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Chenxi Wang, Zhuoyun Yu, Xin Xie, Wuguannan Yao, Runnan Fang, Shuofei Qiao, Kexin Cao, Guozhou Zheng, Xiang Qi, Peng Zhang, and Shumin Deng. SkillX: Automatically constructing skill knowledge bases for agents, 2026a. URL https://arxiv.org/abs/2604.04804.

Dingzirui Wang, Xuanliang Zhang, Keyan Xu, Qingfu Zhu, Wanxiang Che, and Yang Deng. AlignEvoSkill: Towards knowledge-aware and task-aligned agent skill evolution. *arXiv preprint arXiv:2506.23149*, 2025a. doi: 10.48550/arXiv.2506.23149.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Xihang Wang, Zihan Wang, Chengkai Huang, Cao Liu, Ke Zeng, Quan Z. Sheng, and Lina Yao. Purifying multimodal retrieval: Fragment-level evidence selection for RAG. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1880–1891. ACM, 2026b. doi: 10.1145/3805712.3809692.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7190–7204, 2025b. doi: 10.18653/v1/2025.emnlp-main.365.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 63897–63911, 2025c.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

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

Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. Improving probability-based prompt selection through unified evaluation and analysis. *Transactions of the Association for Computational Linguistics*, 12:664–680, 2024. doi: 10.1162/tacl_a_00666.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R. Fung, Hao Peng, and Heng Ji. CRAFT: Customizing LLMs by creating and retrieving from specialized toolsets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=G0vdDSt9XM.

Kun Zeng, Yu Huo, Siyu Zhang, Zi Ye, Yuecheng Zhuo, Haoyue Liu, Yuquan Lu, Junhao Wen, and Xiaoying Tang. Group of skills: Group-structured skill retrieval for agent skill libraries, 2026. URL https://arxiv.org/abs/2605.06978.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

Xinyu Zhao, Zhen Tan, Vaishnav Tadiparthi, Nakul Agarwal, Kwonjoon Lee, Ehsan Moradi Pari, Hossein Nourkhiz Mahjoub, and Tianlong Chen. Generative skill composition for LLM agents, 2026. URL https://arxiv.org/abs/2606.32025.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. Skillrouter: Skill routing for LLM agents at scale, 2026. URL https://arxiv.org/abs/2603.22455.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023. doi: 10.48550/arXiv.2305.17144.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YrycTjllL0.

A Primary Audit and High-Evidence Stress Test

This section reports the fixed-depth GSM8K sweep and the secondary high-evidence audit, followed by the validation and ablation tables referenced in the main text.

Table 4: Leakage-free fixed-k TF-IDF retrieval on GSM8K test (n=1000, Qwen3-8B). Δ is versus k=0 with a paired-task bootstrap 95% CI. No $k\geq1$ improves over abstention.

k	pass@1	Δ vs k=0	95% CI
0	**0.922**	n/a	n/a
1	0.921	-0.001	$[-0.015, +0.013]$
2	0.909	-0.013	$[-0.028, +0.002]$
3	0.916	-0.006	$[-0.021, +0.009]$
5	0.920	-0.002	$[-0.017, +0.013]$

A.1 IG-Ranked and Cross-Domain Stress Test

The primary audit uses gold-free TF-IDF (Table 1). The secondary audit instead retains the three highest-D_i skills from each library and includes in-domain, cross-domain, and mixed pairs. This reference-ranked candidate set tests composition under unusually strong single-skill evidence.

Table 5: Secondary IG-ranked top-3 stress test by pair type (medians, nats, Qwen3-8B). The reference answer is used for candidate ranking, unlike the gold-free primary audit in Table 1.

setting	pair type	n	R	U_1	U_2	S	joint>best
HumanEval	in-in	298	-0.85	$+0.00$	$+0.00$	$+0.83$	0.51
	mixed	897	-0.33	-0.06	$+0.00$	$+0.71$	0.44
	cross-cross	300	-0.22	$+0.00$	$+0.00$	$+0.31$	0.43
GSM8K	in-in	900	$+0.81$	$+0.00$	$+0.00$	-0.32	0.24
	mixed	2700	$+0.39$	$+0.33$	$+0.00$	-0.44	0.13
	cross-cross	900	$+0.35$	$+0.00$	$+0.00$	-0.69	0.03

Reference ranking selects stronger standalone skills, but pair outcomes still vary by domain. On GSM8K, the joint prompt beats the better single in 24% of in-domain pairs and 3% of cross-domain pairs. HumanEval has more positive marginals (Tables 5 and 6).

B Experimental Details

Models and serving. The generator is Qwen3-8B served by vLLM with greedy decoding for free generation and teacher forcing (echoing reference-answer tokens with per-token log-probabilities) for all evidence measurements. No model is fine-tuned; all interventions are prompt-level.

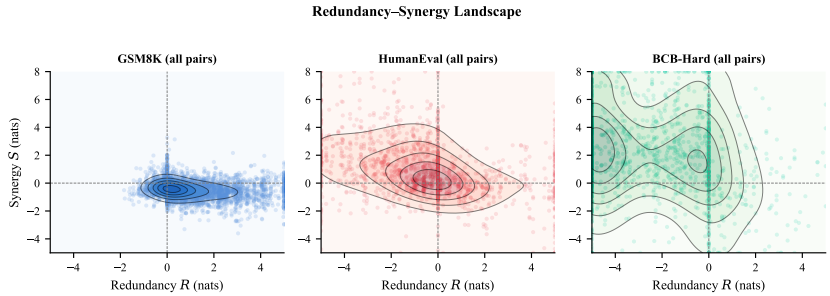


Figure 8: Redundancy-synergy landscape for the secondary IG-ranked stress test. GSM8K concentrates in positive R, negative S; HumanEval and BCB-Hard concentrate in negative R, positive S. The strong GSM8K positive-R concentration is specific to high-D_i candidate ranking and does not appear under the primary TF-IDF protocol.

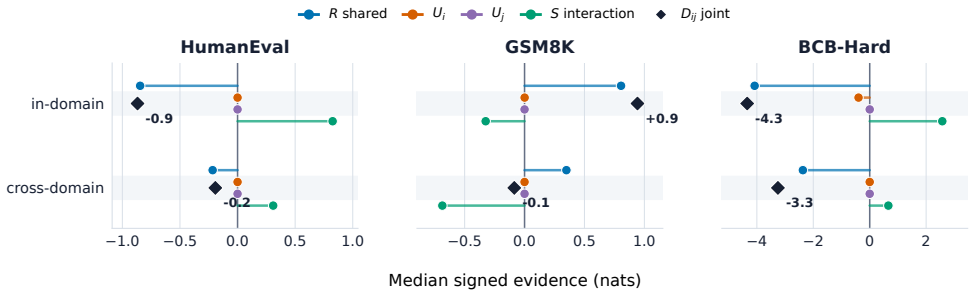


Figure 9: PID atoms in the high-evidence audit. Coloured points show atom medians and black diamonds the median joint contribution. GSM8K combines shared positive evidence with negative interaction; HumanEval and BCB-Hard combine shared negative evidence with positive interaction. Medians need not add even though the decomposition is exact per record; axes differ by panel.

Table 6: Evidence-space marginal of the second skill in the secondary IG-ranked in/cross/mixed stress test. n counts task-pair records. The last column records positive interaction evidence that nevertheless fails to beat the best standalone skill. These are descriptive pair-level rates; behavioural success is evaluated separately.

target / auxiliary library	n	median $\Delta_{a\|b}$	$\Pr(\Delta_{a\|b} > 0)$	$\Pr(S > 0, \Delta_{a\|b} < 0)$
GSM8K / MBPP	4500	-0.59	0.13	0.06
GSM8K / Math500	2996	-0.32	0.25	0.04
HumanEval / MBPP	1495	-0.13	0.45	0.22
HumanEval / BigCodeBench	1495	-0.05	0.48	0.23
BCB-Hard / MBPP	1545	-0.88	0.37	0.39
BCB-Hard / HumanEval	1545	-0.57	0.42	0.33

Table 7: Estimator validation on oracle skills with known information structure (MMLU, Qwen3-8B). Strict Williams-Beer atoms are in bits on tasks where the baseline is wrong; Δ^{beh} is joint accuracy minus the better single. RETRIEVED is the real library.

condition	n	acc_{ab}	Δ^{beh}	R	U_1	S
redundant (ideal 2/0/0)	67	0.81	-0.01	1.11	0.01	0.04
unique (ideal 0/2/0)	67	0.36	-0.45	0.01	1.11	0.01
synergy (ideal 1/0/1)	67	0.12	-0.12	0.08	0.04	0.35
conflict	67	0.27	-0.54	0.42	0.70	0.33
noise (ideal 0/0/0)	67	0.06	$+0.00$	0.01	0.00	0.06
retrieved	67	0.07	-0.01	0.05	0.01	0.08

Table 8: Candidate-set ablation (medians, nats): redundancy and interference persist for random and deliberately dissimilar candidates.

dataset	candidates	n	R	S	D_{ij}	joint>best
HumanEval	retrieved	300	-1.60	$+0.81$	-2.30	0.30
	random	300	-2.00	$+1.33$	-2.33	0.33
	dissimilar	300	-1.17	$+1.42$	-0.98	0.52
GSM8K	retrieved	450	$+0.15$	-0.49	$+0.15$	0.05
	random	448	$+0.10$	-0.30	$+0.30$	0.16
	dissimilar	450	$+0.02$	-0.18	$+0.31$	0.19

Skill libraries. For each benchmark, the library is distilled from the corresponding training split only. An LLM is prompted with a training problem and its ground-truth solution, answer, or canonical code, and asked to extract a reusable skill: generic trigger conditions, a transferable procedure, and anti-overfitting constraints (omitting task-specific constants, names, answer choices, and variable names). Cross-domain libraries are built the same way from a different benchmark (MBPP for the code and math targets, GPQA for MMLU) and applied to the target tasks unchanged. The skill-format check of §5.6 replaces these planning templates with an independently generalised library whose documents emphasise API call patterns and procedural code fragments (BigCodeBench- and MBPP-derived generalised skills), keeping the same retrieval and pair-construction protocol.

Pair and triple construction. For the primary audit, TF-IDF cosine similarity ranks each task's in-domain library after leave-one-task-out filtering. The top-3 candidates yield three pairs per task; only after this gold-free selection do we compute single-skill and pairwise information gain. The high-evidence audit in Appendix A.1 computes D_i for every candidate, retains the top-3 per in-domain and cross-domain library, and forms all in-in (3), cross-cross (3), and mixed (9) pairs. Unless marked as gold-free, the triple, order, cross-domain, and candidate-set analyses use this IG-ranked candidate set. Triples score every non-empty subset of the in-domain top-3, requiring 7 teacher-forced evaluations per task. The original order control re-scores every IG-ranked HumanEval pair with the two skills concatenated in reverse order. The gold-free order audit of Table 12 repeats that swap on every primary TF-IDF pair. Exact-length placebos freeze the same top-1/top-2 pairs and replace the second skill with a whitespace-word-matched irrelevant library skill, a content-free template, or a shuffle of the original top-2 text, then rescore $D_{12} - D_1$ under Qwen3-8B.

Statistical protocol. Pointwise terms are summarised by medians; uncertainty is quantified by resampling independent tasks while preserving all within-task pair records. Sign tests operate on per-task medians. Pair-type comparisons use paired Wilcoxon tests on per-task medians with Holm correction over the family of comparisons. Strict Williams-Beer atoms and outcome-level atoms are reported as observed minus the mean of a 500-replicate permutation null in which Y is shuffled against (X_1, X_2); p-values are the fraction of null replicates reaching the observed value.

Behavioural validation. Each sampled (task, pair) record is run under four free-generation conditions: no skill, skill a only, skill b only, and both. Generations are cached per (task, condition), and the composition marginal is $\mathrm{pass}(ab) - \max\{\mathrm{pass}(a), \mathrm{pass}(b)\}$. The primary protocol uses greedy decoding (T=0, one sample). To reduce Bernoulli noise we additionally re-run HumanEval with T=0.6 and three independent samples per condition, reporting their mean pass rate (Table 18).

C Proof of Proposition 1

(i) Exactness. Immediate from the definitions in Eq. 3: $R + U_i + U_j + S = R + (D_i - R) + (D_j - R) + (D_{ij} - D_i - D_j + R) = D_{ij}$.

(ii) Symmetry. Eq. 2 is invariant under exchanging D_i and D_j: the condition $D_i D_j > 0$ is symmetric, and both $\min(|D_i|, |D_j|)$ and $\mathrm{sgn}(D_i) = \mathrm{sgn}(D_j)$ (on that event) are symmetric.

(iii) Self-redundancy. If $D_i = D_j = d$ with $d \neq 0$ then $D_i D_j = d^2 > 0$, so $R = \mathrm{sgn}(d)\,|d| = d$ and $U_i = U_j = d - d = 0$. If $d = 0$ then $R = 0 = D_i$ and $U_i = U_j = 0$ as well.

(iv) Duplication invariance. Injecting a copy $s_j = s_i$ gives $D_j = D_i$ pointwise, so by (iii) $U_i = U_j = 0$: a duplicate contributes no unique information.

(v) Sign consistency. On the event $D_i D_j > 0$, R carries $\operatorname{sgn}(D_i) = \operatorname{sgn}(D_j)$ by construction; when the two skills disagree in sign there is no shared direction of evidence and $R = 0$. Therefore, $R < 0$ exactly when both skills move the reference answer's log-likelihood downward, i.e. when the shared contribution is negative evidence. □

Proposition 2 (Axiomatic characterisation of signed overlap) *Let $r : \mathbb{R}^2 \to \mathbb{R}$ be a local redundancy functional. Suppose it satisfies* (i) *source symmetry;* (ii) *self-redundancy, $r(c, c) = c$;* (iii) *directional exclusion, $r(a, b) = 0$ when $ab \leq 0$; and* (iv) *weak-source absorption: if $ab > 0$ and $|a| \leq |b|$, all of the weaker collinear contribution is common, so $r(a, b) = a$. Then r is uniquely Eq. 2. Its Möbius atoms also obey source reconstruction, sign-preserving residuals, and $U_a + S = C_{ij} - \max(C_i, C_j)$ independently of r.*

Proof of Proposition 2. Take arbitrary $a, b \in \mathbb{R}$. If $ab \leq 0$, directional exclusion forces $r(a, b) = 0$, matching Eq. 2. If $ab > 0$, either $|a| \leq |b|$ or $|b| \leq |a|$. In the first case weak-source absorption gives $r(a, b) = a = \operatorname{sgn}(a) \min(|a|, |b|)$. In the second, symmetry followed by absorption gives $r(a, b) = r(b, a) = b = \operatorname{sgn}(a) \min(|a|, |b|)$. The axioms therefore fix every input, is fixed by the axioms and Eq. 2 is unique. Conversely, direct inspection shows that Eq. 2 satisfies all four axioms.

Source reconstruction follows from $U_i = C_i - R$. For same-sign inputs, R has their common sign and no larger magnitude than either source, so each nonzero residual retains that sign. For opposite-sign inputs, $R = 0$, so each residual equals its source. Finally, let b index the better singleton and a the other. Möbius inversion gives

$$U_a + S = (C_a - R) + (C_{ij} - C_a - C_b + R) = C_{ij} - C_b = C_{ij} - \max(C_i, C_j),$$

so the addition value is invariant to the chosen redundancy functional. □

D Oracle Skill Constructions

Oracle skills are synthesised per MMLU task from the gold option g and distractors d_1, d_2, d_3:

- REDUNDANT: both skills are short notes asserting that the correct answer is g (paraphrased differently to avoid string duplication). Ideal atoms $R = H(Y), U_1 = U_2 = S = 0$.
- UNIQUE: skill 1 asserts g; skill 2 is an unrelated true fact of matched length. Ideal $U_1 = H(Y)$, $R = U_2 = S = 0$.
- SYNERGY: skill 1 asserts "the answer is g or d_1"; skill 2 asserts "the answer is g or d_2". Neither determines Y; their intersection does. With four balanced options, ideal $R = S = 1$ bit.
- CONFLICT: skill 1 asserts g; skill 2 asserts d_1.
- NOISE: two unrelated facts. Ideal $R = U_1 = U_2 = S = 0$.

Because the generator answers many MMLU items correctly without help, all atoms in Table 7 are computed on the subset of tasks the no-skill baseline gets wrong ($n = 67$), where the injected hints rather than prior knowledge drive the prediction; results over all tasks show the same ordering with attenuated magnitudes. Figure 3 visualises the recovered atoms against the designed structure.

E Full Strict Decomposition by Pair Type

Table 9 breaks the strict Williams-Beer decomposition of §5.5 down by pair type. The redundancy-dominant ordering holds within every group; in-domain pairs carry the largest synergistic share and cross-domain pairs the largest redundant share.

F Prompt-Construction Controls

F.1 Injection Order

Although the aggregate signed order effect is centred near zero (mean $+0.020$ nats; task-pair bootstrap 95% CI $[-0.059, 0.099]$), its median absolute magnitude is 0.721 nats. Reversing the two

Table 9: Strict Williams-Beer PID on multiple-choice answers (MMLU, bits) by pair type. *null* is the mean value under 500 label permutations; *corrected* subtracts it.

group	atom	observed	null	corrected	p
cross-cross	$I(Y; X_1, X_2)$	0.348	0.069	0.279	0.002
	R	0.237	0.010	0.227	0.002
	U_1	0.006	0.005	0.001	0.295
	U_2	0.027	0.005	0.022	0.002
	S	0.078	0.050	0.028	0.010
in-in	$I(Y; X_1, X_2)$	0.277	0.070	0.207	0.002
	R	0.147	0.009	0.138	0.002
	U_1	0.044	0.006	0.038	0.002
	U_2	0.004	0.006	−0.001	0.521
	S	0.081	0.049	0.032	0.006
mixed	$I(Y; X_1, X_2)$	0.325	0.024	0.301	0.002
	R	0.163	0.003	0.160	0.002
	U_1	0.000	0.002	−0.002	0.936
	U_2	0.088	0.002	0.087	0.002
	S	0.074	0.017	0.056	0.002

Table 10: Outcome-level Williams–Beer PID on HumanEval ($n = 180$; bits). The target is joint-run success; the sources are the two single-skill success indicators. Corrected values subtract the mean of 1,000 label permutations.

atom	observed	null	corrected	p
$I(Y; X_1, X_2)$	0.598	0.012	0.585	0.001
R	0.455	0.002	0.453	0.001
U_1	0.000	0.002	−0.002	0.511
U_2	0.033	0.002	0.031	0.001
S	0.110	0.007	0.103	0.001

Table 11: Injection-order control on HumanEval. $\Delta = D_{ij}^{(a \text{ first})} - D_{ij}^{(b \text{ first})}$ on identical skill pairs; a median far from zero would indicate that asymmetries attributed to the skills are prompt-position effects.

pair type	n	median Δ	frac $\lvert\Delta\rvert > 0.5$
cross-cross	300	+0.061	0.62
in-in	298	+0.150	0.64
mixed	895	−0.084	0.62
all	1493	−0.017	0.62

skills changes the sign of the second-skill marginal on 26.3% of pairs and the sign of S on 22.4%; averaging the two orders changes 13.9% and 11.7%, respectively, relative to the original one-order labels. Order has little aggregate effect but can change individual pair labels.

Table 12 repeats the swap on the primary gold-free TF-IDF pairs. Across all seven datasets the mean order effect remains consistent with zero, while the median better-single marginal stays negative in both concatenations.

Table 12: Gold-free injection-order control. Mean $\Delta_{\text{order}} = D_{ab} - D_{ba}$ with task-bootstrap 95% CIs; median $\Delta_{a|b}$ in the original and reversed concatenations; Spearman ρ of pair scores across orders; fraction of pairs whose marginal sign flips.

dataset	n	mean Δ_{order}	95% CI	med. Δ orig. / rev.	ρ	flip
GSM8K	300	+0.041	[−0.040, +0.118]	−0.36 / −0.37	0.865	0.25
Math500	300	−0.009	[−0.074, +0.062]	−0.53 / −0.50	0.895	0.15
GPQA	300	+0.010	[−0.039, +0.058]	−0.24 / −0.25	0.877	0.24
MMLU	300	−0.078	[−0.206, +0.052]	−0.52 / −0.51	0.912	0.21
HumanEval	192	+0.232	[−0.068, +0.541]	−0.65 / −0.76	0.899	0.19
MBPP	300	+0.288	[−0.165, +0.748]	−2.96 / −2.87	0.948	0.17
BCB-Hard	300	−0.195	[−0.571, +0.178]	−1.00 / −0.99	0.911	0.25

F.2 Exact-Length Placebos

Table 13: Exact-length second-slot controls on fixed gold-free top-1/top-2 pairs (Qwen3-8B). Entries are mean $D_{12} - D_1$; controls replace the second skill with a matched irrelevant skill, a content-free template, or shuffled top-2 text. The final column gives original minus irrelevant with a task-bootstrap 95% CI.

dataset	original	irrelevant	template	shuffled	orig.−irr. [CI]
GSM8K (n=100)	+0.23	+0.41	+0.47	+0.86	−0.19 [−0.44, +0.20]
MMLU (n=100)	+0.15	+0.02	−1.12	−0.47	+0.13 [−0.05, +0.31]
HumanEval (n=64)	+0.70	+2.31	+2.75	+2.99	−1.62 [−2.26, −0.97]

On GSM8K the shuffled control exceeds the retrieved skill by 0.63 nats (CI $[0.51, 0.76]$); on HumanEval all three families do (shuffled difference 2.29, CI $[1.67, 2.97]$). MMLU is the exception among content-free templates: empty or shuffled text is harmful, so the retrieved pair beats those controls but not an equal-length irrelevant skill. A positive $D_{12} - D_1$ alone is therefore insufficient evidence of semantic complementarity.

G Extensions of the Pairwise Audit

G.1 Injection Depth and Higher-Order Marginals

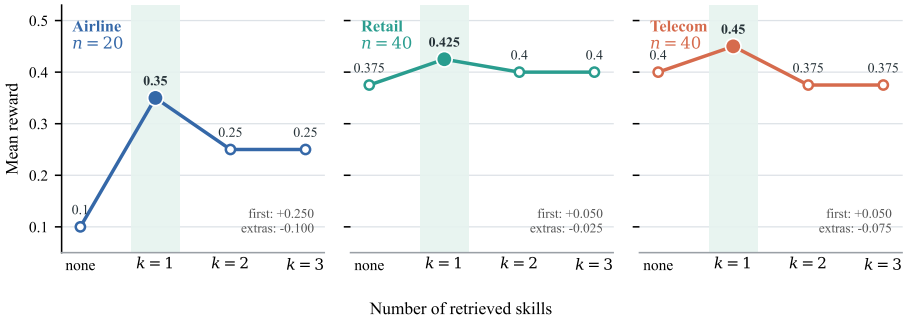


Figure 10: τ^2-Bench injection-depth sweep with the official SkillX pipeline. Across airline, retail, and telecom, k=1 improves over no-skill while $k \in \{2, 3\}$ reduces mean reward relative to k=1.

Table 14: Triple-skill residuals in the high-evidence audit.

dataset	n	median S_3^{marg}	frac. > 0	median IE_3
HumanEval	100	−0.45	0.31	−1.99
GSM8K	150	−0.23	0.29	+1.39
BCB-Hard	103	−2.98	0.12	−10.05

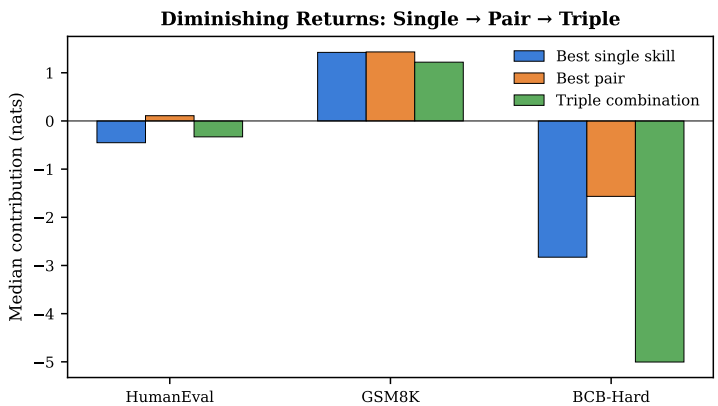


Figure 11: Diminishing returns from single to pair to triple skill injection (median contribution in nats).

Table 15: BigCodeBench-Hard decomposition with MBPP cross-domain skills (medians, nats).

pair type	n	R	S	D_{ij}	joint>best
in-in	309	-4.08	$+2.57$	-4.34	0.48
mixed	927	-2.26	$+2.60$	-3.41	0.32
cross-cross	309	-2.37	$+0.66$	-3.25	0.40



Figure 12: Median PID atoms across six library-transfer settings. Same-type transfer does not create complementarity.

Table 16: Same-type skill transfer (medians, nats). Math→math has positive redundancy; code→code has negative redundancy and positive interaction.

task / cross library	pair type	n	R	S	D_{ij}
GSM8K / Math500	in-in	600	$+0.69$	-0.32	$+0.88$
	mixed	1800	$+0.62$	-0.21	$+1.06$
	cross-cross	600	$+0.72$	-0.35	$+0.73$
BCB-Hard / HumanEval	in-in	309	-2.73	$+2.72$	-2.29
	mixed	927	-1.92	$+2.07$	-2.90
	cross-cross	309	-2.38	$+0.79$	-2.73
HumanEval / BCB-Hard	in-in	300	-0.84	$+0.73$	-0.96
	mixed	900	-0.10	$+0.81$	-0.61
	cross-cross	300	-0.05	$+0.40$	-0.27

G.2 Transfer Between Related Domains

G.3 Constructed Complement

Table 17: Illustrative GSM8K quantity-formula construction (§5.6). Evidence-space atoms are medians (nats).

setting	n	S	D_{both}	joint>best
GSM8K qty×formula	105	$+9.82$	$+69.0$	0.84

G.4 Behavioural and Skill-Format Checks

H Distribution-Space Diagnostics

Following decomposition-inspired probes for multimodal models (Liang et al., 2023a), we compute four features directly on the answer distributions P_1, P_2, P_{12}:

$$\begin{aligned} R_{\text{flow}} &= 1 - \text{JSD}(P_1 \| P_2), & U_1 &= D_{\text{KL}}(P_{12} \| P_2), \\ S_{\text{ent}} &= \tfrac{1}{2}(H(P_1) + H(P_2)) - H(P_{12}), & U_2 &= D_{\text{KL}}(P_{12} \| P_1). \end{aligned} \tag{7}$$

Table 18: Left: multi-sample outcome stratification on HumanEval (T=0.6, 3×). Right: skill-format check with generalised BigCodeBench/MBPP documents (medians, nats).

Multi-sample outcomes				*Generalised skill format*				
bucket	n	S	ab	pair type	n	R	S	D_{ij}
interference	30	-4.40	0.50	in-in	300	-0.65	$+0.41$	-0.76
control	30	$+0.00$	0.71	mixed	900	-0.58	$+0.87$	-0.93
synergy	30	$+7.59$	0.39	cross-cross	300	-1.10	$+0.81$	-0.99

Here R_{flow} measures agreement between P_1 and P_2 without reference to Y, while S_{ent} measures confidence sharpening. Neither quantity distinguishes confident errors, so we use them as distributional summaries alongside Eqs. 3 and 6.

I Sensitivity to the Redundancy Definition

The pointwise decomposition requires an independent redundancy definition (Eq. 2); both unique atoms and $S = D_{ij} - D_i - D_j + R$ inherit it. Of the four conventions below, signed overlap satisfies all axioms of Proposition 2. Positive-only discards shared negative evidence, absolute overlap reverses its sign, and $R \equiv 0$ removes the shared atom. These alternatives show how the allocation changes when individual axioms are relaxed.

Unique-atom sensitivity. Table 19 reports median unique atoms under each definition. Exact zero medians depend on the overlap convention; with $R \equiv 0$, the unique atoms reduce to the raw singleton contributions D_i, D_j and are generally nonzero.

Table 19: Primary gold-free audit: median U_i/U_j (nats) under four redundancy conventions. Exact zero medians vary with the redundancy definition.

dataset	signed overlap	positive-only	absolute overlap	$R \equiv 0$
GSM8K	$+0.00/+0.00$	$-0.09/-0.16$	$-0.17/-0.32$	$-0.09/-0.16$
Math500	$+0.00/+0.00$	$-0.32/-0.36$	$-0.52/-0.56$	$-0.32/-0.36$
GPQA	$+0.00/+0.00$	$+0.00/+0.00$	$+0.00/+0.00$	$+0.23/+0.35$
MMLU	$+0.00/+0.00$	$-1.17/-1.48$	$-2.09/-2.50$	$-1.17/-1.48$
HumanEval	$+0.00/-0.30$	$-6.72/-6.84$	$-11.99/-11.74$	$-6.72/-6.84$
MBPP	$+0.00/+0.00$	$-6.49/-5.28$	$-11.54/-9.09$	$-6.49/-5.28$
BCB-Hard	$-0.18/+0.00$	$-7.29/-7.20$	$-12.01/-12.59$	$-7.29/-7.20$

Two statistics computed before choosing R are reported in Table 20: the fraction of records in which D_i, D_j have the same direction and, conditional on alignment, their balance $q = \min(|D_i|, |D_j|)/\max(|D_i|, |D_j|)$. Across datasets, 73% to 95% of singleton effects align and the weaker contribution retains 51% to 71% of the stronger one's magnitude. These observations show substantial common directional evidence. Separate strict checks assign 62% of corrected MMLU answer information and 77% of corrected HumanEval joint-success information to redundancy (Tables 3 and 10).

Table 20: Definition-free singleton agreement in the primary audit. q is computed only on same-direction records and uses no redundancy functional.

dataset	n	$\Pr(D_i D_j > 0)$	median $q \mid D_i D_j > 0$
GSM8K	300	0.73	0.51
Math500	300	0.74	0.55
GPQA	300	0.81	0.55
MMLU	300	0.83	0.64
HumanEval	192	0.95	0.71
MBPP	300	0.83	0.57
BCB-Hard	300	0.80	0.64

Interaction sensitivity. Table 21 and Figure 13 recompute median S under four choices: signed overlap (Eq. 2), positive-only, absolute, and $R \equiv 0$. Across all definitions, the qualitative conclusions of §5.1 are unchanged: median S is negative for every GSM8K pair type and positive for every HumanEval pair type.

On code benchmarks, S remains positive under all four definitions. Making R less negative enlarges S mechanically while leaving D_{ij} unchanged.

Table 21: Median pointwise synergy S (nats) / fraction $S > 0$ under four redundancy definitions (Qwen3-8B).

dataset	pair type	signed	positive-only	absolute	$R \equiv 0$
HumanEval	in-in	$+0.83$ / 0.69	$+2.04$ / 0.82	$+2.81$ / 0.85	$+1.46$ / 0.70
	mixed	$+0.71$ / 0.69	$+1.55$ / 0.78	$+2.20$ / 0.80	$+1.29$ / 0.67
	cross-cross	$+0.31$ / 0.61	$+1.09$ / 0.74	$+1.75$ / 0.77	$+0.90$ / 0.65
GSM8K	in-in	-0.32 / 0.28	-0.30 / 0.29	-0.28 / 0.32	-1.20 / 0.11
	mixed	-0.44 / 0.21	-0.42 / 0.24	-0.35 / 0.29	-0.98 / 0.17
	cross-cross	-0.69 / 0.05	-0.61 / 0.17	-0.58 / 0.23	-1.14 / 0.16
BCB-Hard	in-in	$+2.57$ / 0.79	$+7.52$ / 0.91	$+11.25$ / 0.93	$+7.52$ / 0.89
	mixed	$+2.60$ / 0.80	$+5.57$ / 0.88	$+8.44$ / 0.91	$+5.57$ / 0.86
	cross-cross	$+0.66$ / 0.62	$+3.73$ / 0.82	$+6.32$ / 0.86	$+3.60$ / 0.79

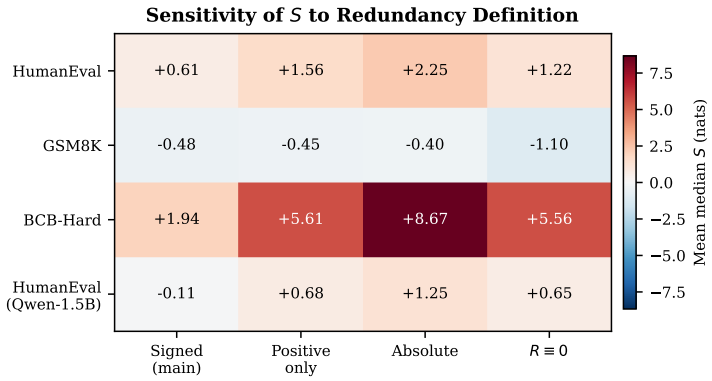


Figure 13: Mean median S under four redundancy definitions.

J Similarity and Marginal Value

We compare skill-pair Jaccard similarity with redundancy, the evidence marginal, and the free-generation marginal (Table 22). The correlations with the evidence marginal are -0.12 on GSM8K and -0.06 on HumanEval; AUROCs for a positive marginal are 0.40 and 0.47. Figure 14 shows the pair-level relationship. Among GSM8K pairs with high overlap and low uniqueness, this regime contains $2984/4500$ pairs, and the joint beats the better single in 14% of those pairs.

Table 22: Skill-pair Jaccard similarity versus PID atoms and marginal value after the best single skill. Correlations are Spearman ρ. AUROC is for predicting $D_{ij} - D_{\text{best}} > 0$ from similarity alone.

dataset	n	ρ(sim,R)	ρ(sim,marginal)	ρ(sim,beh. marginal)	AUROC
GSM8K	4500	$+0.17$	-0.12	-0.11 (n=180)	0.40
HumanEval	1495	-0.03	-0.06	-0.02 (n=181)	0.47

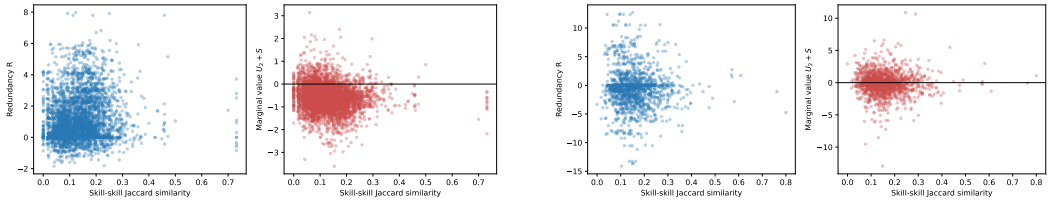


Figure 14: Skill-pair Jaccard similarity versus redundancy R (left panels within each plot) and evidence-space marginal value after the best single skill (right panels). **Left pair:** GSM8K (n=4500). **Right pair:** HumanEval (n=1495). Similar second skills do not systematically raise marginal value.

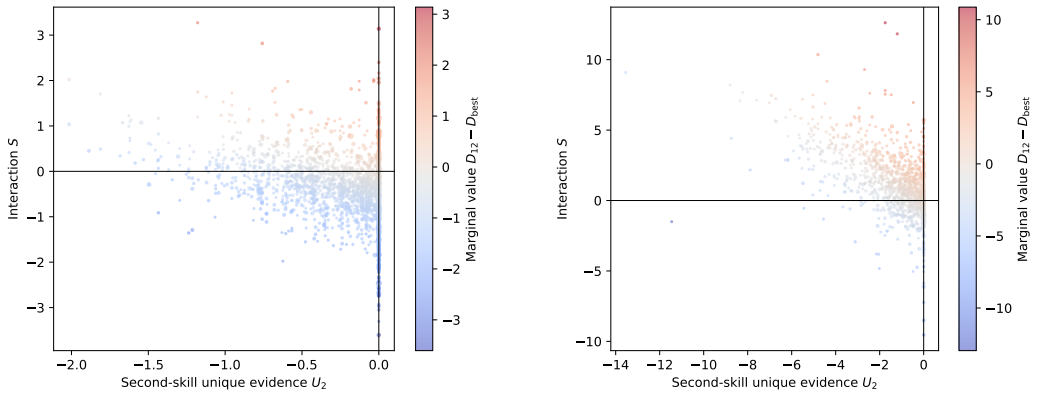


Figure 15: Signed-overlap second-skill unique evidence U_2 versus interaction S, coloured by marginal value after the best single skill (marker size $\propto$ second skill length). **Left:** GSM8K. **Right:** HumanEval. Positive marginal cases concentrate where S is positive; high-R/near-zero-U_2 regions dominate under this allocation and rarely help.

K Interaction Distributions

K.1 Synergy Distribution Across Datasets

Figure 16 shows the full distribution of pointwise synergy S across three datasets and three pair types (in-domain, mixed, cross-domain). On GSM8K, synergy is predominantly negative across all pair types, with cross-domain pairs showing the most negative medians. On HumanEval, synergy is positive for in-domain pairs but shifts toward zero for cross-domain pairs. BCB-Hard shows broadly positive evidence S, driven by large IG magnitudes.

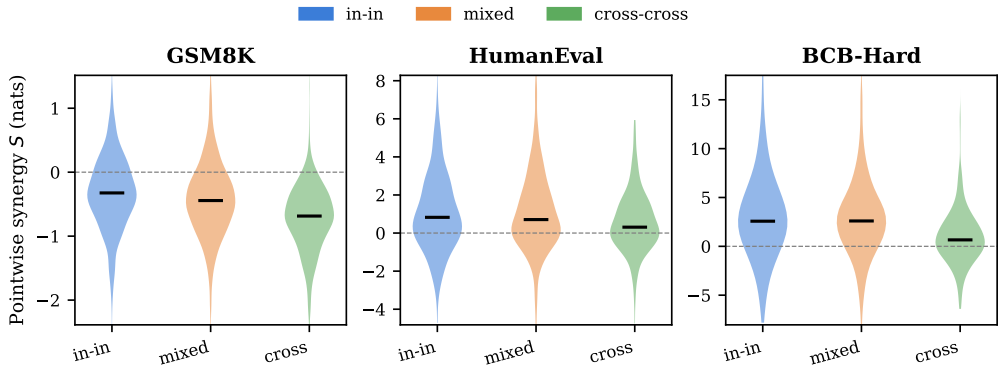


Figure 16: Violin plots of pointwise synergy S by pair type across three datasets. Black bars indicate medians. Synergy sign and spread vary across domains: GSM8K is predominantly negative, HumanEval is mixed-positive, and BCB-Hard is broadly positive.

L Hierarchical Agent Skills (ALFWorld)

We apply the signed pointwise decomposition (Eqs. 2 and 3) to a hierarchical AlfWorld skill bank (Shridhar et al., 2021) with the frozen Qwen3-8B generator. Skills are written at two granularities (task principles and linked step principles). Related candidates are selected by gold-free TF-IDF (leave-one-task-out top-3 task skills; rank-aligned task-step pairs), then scored under teacher forcing of an expert plan.

Same-granularity task-task pairs are redundancy-dominated ($R > 0$ in 90.9% of pairs) and usually fail to beat the better single skill (joint>best $= 42.3\%$; mean marginal -1.96 nats). Under teacher forcing of the expert plan, task-step pairs have a positive evidence marginal; live rollouts for the same pair type have a negative accuracy marginal.

Table 23: AlfWorld hierarchical evidence PID under TF-IDF top-3 selection (Qwen3-8B, 300 tasks). "marg." is the mean evidence marginal after the best single skill.

pair type	n	marg.	joint>best	med. R	med. S	$R > 0$
task-task	900	-1.96	0.423	$+20.7$	-1.22	0.909
task-step	900	$+6.29$	0.740	$+19.0$	$+7.18$	0.904
all	1800	$+2.16$	0.582	$+19.7$	$+2.71$	0.907

Behavioural rollouts. We run live episodes with a Qwen3-8B agent (25-step cap, ten rollouts per condition, temperature 0.3) on 50 task-task and 50 task-step pairs (Table 24; Figure 17). The task-task composition marginal is -0.088, versus -0.034 for task-step pairs.

Table 24: AlfWorld interactive four-condition rollouts (Qwen3-8B, 10 rollouts per condition, 25-step cap, temperature 0.3). *empty*/*best*/*joint* are mean success rates with no skill, the better single skill, and both skills; $\bar{\Delta}^{\text{beh}}$ and joint>best are relative to the better single skill.

pair set	n	empty	best	joint	$\bar{\Delta}^{\text{beh}}$	joint>best
task-task	50	0.126	0.486	0.398	-0.088	0.060
task-step	50	0.124	0.394	0.360	-0.034	0.060
all	100	0.125	0.440	0.379	-0.061	0.060

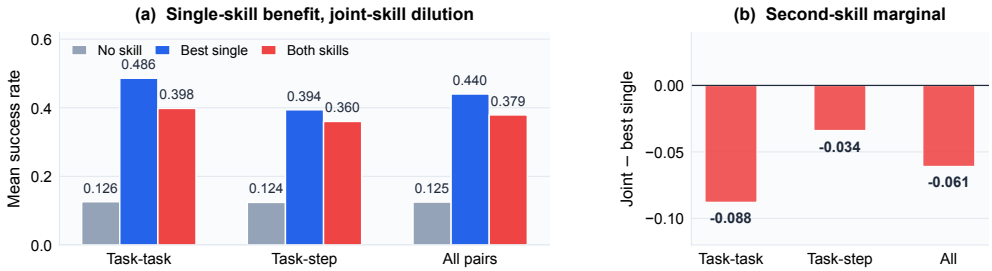


Figure 17: Repeated AlfWorld rollouts: a single retrieved skill markedly improves success over no skill, whereas appending a second related skill reduces mean success for both task-task and task-step pairs. Bars show mean success across 50 pairs per type and 10 rollouts per condition.