Divergent Thinking: Escape the Homogeneity Trap in Generative Commonsense Reasoning

Yiheng Tao[1,3]*, Kaiwen Cheng[1,4]*, Zhiwei Nie[1,3], Chang Liu[5]†, Jie Chen[2,3,1,4]†

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2]School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, China
[3]Pengcheng Laboratory, Shenzhen, China
[4]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[5]Department of Automation and BNRist, Tsinghua University, Beijing, China

Abstract

Generative commonsense reasoning (GCR) requires models to synthesize coherent narratives that simultaneously satisfy lexical constraints and commonsense logic. Although ensemble-based LLM strategies are widely adopted to alleviate the fragility of single-chain reasoning, we uncover a counterintuitive homogeneity trap in GCR. Specifically, we observe that increasing the number of reasoning chains can degrade performance, as the generated chains tend to collapse into a narrow semantic region, thereby reinforcing shared biases rather than providing complementary evidence. We posit that escaping this trap requires fundamentally broadening semantic coverage via heterogeneous sources. Our investigation into the nature of diversity reveals that deep semantic diversity, rather than surface-level lexical variation, is the decisive prerequisite for effective integration. Motivated by this insight, we propose an Explore-then-Integrate framework, in which high–semantic-entropy explorers capture diverse concept bindings, and a powerful integrator performs compositional synthesis to merge valid fragments into coherent narratives. Crucially, to ensure that the observed performance gains arise from accurate logical composition rather than trivial best-candidate selection, we introduce a provenance-aware evaluation suite that explicitly quantifies the heterogeneous origins of synthesized outputs. Extensive experiments on multiple benchmarks demonstrate the consistent superiority of our approach across a range of metrics. Notably, our method achieves over **10%** improvement in overall accuracy on NoRa and in SPICE score on CommonGen-Lite.

1 Introduction

Generative commonsense reasoning (GCR) serves as a critical testbed for evaluating the compositional generalization and reasoning capabilities of large language models (LLMs) (Lin et al., 2020; Kojima et al., 2022). Functioning as a complex constraint satisfaction problem, GCR demands the synthesis of fluent narratives that strictly adhere to lexical constraints and commonsense logic, while maintaining semantic diversity (Yu et al., 2022; Zhang et al., 2025). Although Chain-of-Thought (CoT) prompting (Wei et al., 2022) has demonstrated remarkable success in various reasoning

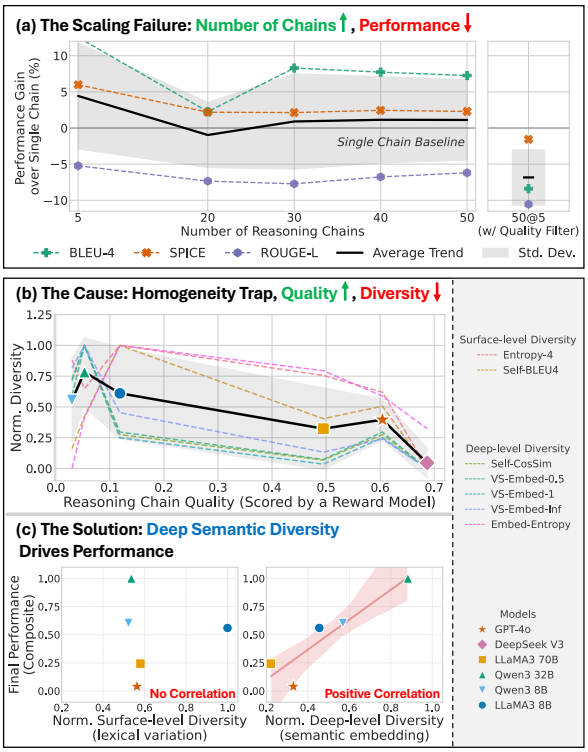

Figure 1: Escaping the Homogeneity Trap. (a) **Scaling Failure.** Conventional ensemble strategies exhibit performance degradation as the number of reasoning chains increases, even with quality filtering. (b) **The Homogeneity Trap.** An evaluation of 50 sampled chains across 7 metrics shows that as models become more capable (e.g., GPT-4o vs. Qwen-8B), they exhibit pronounced mode collapse. (c) **Effective Diversity.** Surface-level diversity correlates poorly with final performance, whereas deep-level diversity exhibits a strong positive correlation, proving decisive for integration.

*Equal contribution.
†Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18395–18410
July 2-7, 2026

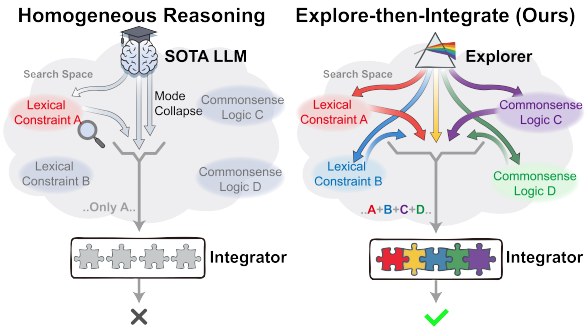

Figure 2: A comparison between mode-collapsed homogeneous reasoning and our Explore-then-Integrate framework.

tasks (Kojima et al., 2022; Zhou et al., 2023; Zhang et al., 2023b), it exhibits significant limitations due to "Constraint Attention Drop" (Li et al., 2025). Specifically, the generation of explicit reasoning steps often unintentionally diverts the model's focus from critical constraint tokens, resulting in the hallucination of irrational relationships to force-fit concepts (Zhang et al., 2023a).

To mitigate single-chain fragility, ensemble strategies, such as voting-based Self-Consistency (SC) (Wang et al., 2023) and meta-reasoner-based Multi-Chain Reasoning (MCR) (Yoran et al., 2023), have become standard practice, typically relying on repeated sampling from a single powerful model. However, contrary to previous findings, we observe a scaling failure when applying these methods to GCR: increasing the sampling budget yields diminishing returns or even performance degradation, despite the application of top-quality candidate filtering, as shown in Fig. 1 (a). We attribute this phenomenon to the **Homogeneity Trap**. As validated in Fig. 1 (b), state-of-the-art LLMs exhibit a distinct inverse correlation where higher reasoning quality comes at the cost of sharply reduced diversity. Consequently, generated chains cluster within a narrow semantic band and share identical systematic biases (Wu et al., 2025; Jiang et al., 2025). In this regime, the "wisdom of the crowd" collapses; high redundancy merely amplifies shared biases and hampers the composition ability of the integrators (e.g., meta-reasoners) to synthesize concepts into a reliable final output.

To escape the homogeneity trap, we propose expanding the search space via heterogeneous sources, grounded in the cognitive intuition that *compositional synthesis is more tractable than ab-initio generation* (Russin et al., 2025; Sinha et al., 2024). While homogeneous inputs force integrators to reconstruct missing logic from scratch, a diverse candidate pool enables fragment-level composition, transforming complex generation into simpler recognition and merging. However, distinct input sources do not inherently guarantee better integration. We investigate the prerequisites for effective integration and reveal a critical distinction: while surface-level diversity (lexical variation) yields limited utility, deep-level diversity (semantic embedding) is the decisive driver (Fig. 1 c). Guided by this, we introduce **Explore-then-Integrate**. We instantiate this by deploying a high-semantic-entropy explorer to capture imperfect yet complementary concept bindings, enabling a powerful integrator to synthesize these diverse fragments into a coherent, constraint-satisfying narrative (Fig. 2).

However, strictly validating our framework requires opening the black box of integration. We propose a fine-grained **Provenance-aware Evaluation Suite**. By quantifying semantic provenance, we verify that our method performs granular integration—weaving fragments from distinct sources—rather than trivial best-candidate selection. Crucially, this analysis also dispels concerns about high-entropy noise (Farquhar et al., 2024): we reveal that heterogeneity explicitly exposes conflicts, thereby simplifying arbitration compared to the deceptive consensus of homogeneous chains. Furthermore, to address the efficiency overhead of multiple LLM calls, we propose **Distilled Integration**, demonstrating that this high-level fusion logic can be internalized into a compact student model for efficient, standalone integration.

We conduct extensive experiments on several public benchmarks. Experimental results demonstrate that our framework can escapes the homogeneity trap to a large extent, delivering over 10% SPICE gains on Commongen-Lite and achieving a state-of-the-art 90.2% accuracy on NoRa. Our distilled student model successfully internalizes the complex integration policy and exhibits remarkable cross-dataset generalization on DimonGen.

2 Related Work

2.1 Generative Commonsense Reasoning

LLMs often lack robust commonsense capabilities and can generate sentences that contradict basic commonsense knowledge. This limitation is partly due to the well-recognized reporting bias (Gordon and Van Durme, 2013), where the amount of explicitly recorded commonsense information in text is far less than its prevalence in the real world (Havasi

et al., 2007). Generative Commonsense Reasoning (GCR) requires composing coherent narratives under rigid concept constraints (Lin et al., 2020). Early approaches primarily focused on retrieving external commonsense knowledge to augment generation (Liu et al., 2021; Guan et al., 2020; Li et al., 2021; He et al., 2022; Liu et al., 2022, 2023; Cui et al., 2024). These methods depend heavily on task-specific supervision and lack open-ended generalization. With the advent of LLMs, recent benchmarks such as CommonGen-Lite (Lin et al., 2020) and NoRa (Zhou et al., 2024) have shifted focus to prompting strategies, most of which rely on single-chain reasoning and struggle to balance constraint coverage with commonsense logic.

2.2 Chain-of-Thought and LLM Ensemble

While Chain-of-Thought (CoT) prompting enhances reasoning, it frequently suffers from "constraint attention drop" (Li et al., 2025) in constraint-heavy tasks, sacrificing hard constraints for partial fluency. Post-inference LLM ensembles (Chen et al., 2025; Lu et al., 2024) attempt to mitigate this via selection-based (Wang et al., 2023; Li et al., 2024; Guha et al., 2024) or selection-then-regeneration frameworks (Jiang et al., 2023; Yoran et al., 2023; Lv et al., 2024). However, limitations persist: selection prioritizes high-probability consensus over necessary semantic diversity, while regeneration often functions merely as selection rather than true fragment-level composition. In contrast, our Explore-then-Integrate framework couples high-entropy exploration with prudent integration for true fragment-level synthesis.

3 Methodology

We introduce Explore-then-Integrate, grounded in the intuition that *compositional synthesis is more tractable than ab-initio generation*. Unlocking this potential requires a candidate pool exhibiting deep semantic diversity. Guided by this, our workflow (Fig. 3) proceeds through three phases: (1) Exploration, deploying uninhibited models to maximize semantic entropy and capture diverse concept bindings; (2) Integration, where a powerful reasoner performs fragment-level synthesis to merge coherent logic; and (3) Distilled Integration, internalizing this fusion capability into a student model for efficient, standalone inference.

3.1 Problem Formulation

GCR is formulated as a conditional generation task (Lin et al., 2020) mapping an unordered concept set $\mathcal{C} = \{c_1, \ldots, c_N\}$ of size N to an output sentence y. The objective is to maximize $p(y|\mathcal{C})$ subject to two constraints: (1) lexical coverage, requiring all concepts in $\mathcal{C}$ to appear with appropriate morphological inflection, and (2) rationality, ensuring the narrative adheres to commonsense logic. To explicitly model the reasoning process, we decompose the output into a reasoning chain r (marked by `<think>` tags (Wei et al., 2022)) and a final response s, i.e., $y = (r, s)$.

3.2 Phase I: Exploration

The objective of this phase is to construct a high-entropy yet valid candidate pool $\mathcal{H}$. We aim to maximize deep semantic diversity, capturing the distributional tail of valid reasoning paths to cover the semantic search space of the concept set $\mathcal{C}$.

3.2.1 Heterogeneous Sampling

To escape the homogeneity trap, we designate a high-semantic-entropy explorer $\mathcal{M}_{\text{exp}}$ (e.g., Qwen-32B (Bai et al., 2023)). Unlike the counterparts that often converge to narrow, safety-aligned semantic zones, $\mathcal{M}_{\text{exp}}$ maintains a more uninhibited probability distribution. Using the exploration instruction $\mathcal{I}_{\text{exp}}$ (A.1), we sample M candidates $\mathcal{H} = \{h_1, \ldots, h_M\}$ with a high temperature T:

$$h_i \sim p_{\mathcal{M}_{\text{exp}}}(h|\mathcal{C}, \mathcal{I}_{\text{exp}}; T) \quad (1)$$

This strategy encourages the model to explore the distributional tail, capturing varied concept bindings (e.g., polysemous interpretations) that aligned models often suppress, thereby providing diverse logical fragments for downstream synthesis.

3.2.2 Filtering via Pretrained Reward Model

To purge noise from high-entropy sampling without sacrificing diversity, we employ a Pretrained Reward Model (PRM) as a validity filter. Crucially, unlike likelihood-based selection which reinforces homogeneity by favoring commonality, our strategy strictly targets irrational hallucinations, preserving the long-tail of diverse but valid reasoning. We formalize the evaluation of each candidate $h \in \mathcal{H}$ as a multi-dimensional vector $\mathbf{v}(h) \in \mathbb{R}^4$, spanning four axes: Coherence (v_{coh}), Length (v_{len}), Lexical Coverage (v_{cov}), and Information Density (v_{den}). The scoring function $f(h)$

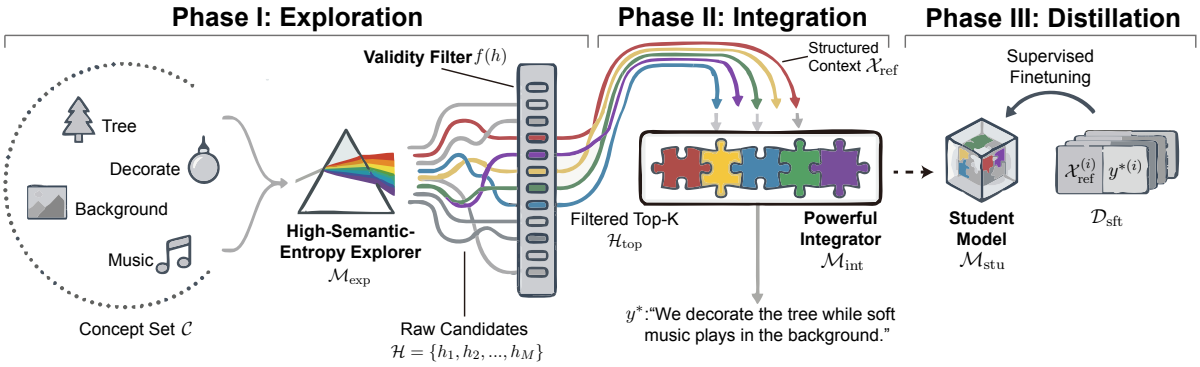

Figure 3: Illustration of our proposed Explore-then-Integrate Framework.

is defined as the projection (A.2):

$$f(h) = \mathbf{w}^\intercal \cdot \mathbf{v}(h) \tag{2}$$

where $\mathbf{w}$ is a preference vector calibrated to penalize violations of hard constraints (e.g., missing concepts) while accommodating the diverse, high-entropy expressions. We select the top-K candidates $\mathcal{H}_{\text{top}} = \text{TopK}_{h \in \mathcal{H}} f(h)$. This ensures the downstream integrator is supplied with a context set that is logically rigorous yet semantically heterogeneous.

3.3 Phase II: Integration

In this phase, we deploy a powerful integrator $\mathcal{M}_{\text{int}}$ (e.g., GPT-4o (OpenAI, 2023)) to perform fragment-level synthesis. Grounded in our core intuition, $\mathcal{M}_{\text{int}}$ is tasked not with creating narratives from scratch, but with decomposing and amalgamating valid logical fragments from the filtered heterogeneous set $\mathcal{H}_{\text{top}}$. We first construct a structured context $\mathcal{X}_{\text{ref}}$ by injecting the diverse reasoning paths from $\mathcal{H}_{\text{top}}$ as distinct reference attempts:

$$\mathcal{X}_{\text{ref}} = \bigoplus_{h \in \mathcal{H}_{\text{top}}} (\texttt{[Attempt ID]} : h) \tag{3}$$

where $\bigoplus$ denotes concatenation. This layout explicitly exposes semantic contrasts, transforming the task into a discriminative puzzle where the model identifies complementary strengths across inputs. To prevent superficial copying and enforce true composition, the fusion instruction $\mathcal{I}_{\text{fuse}}$ (A.3) compels $\mathcal{M}_{\text{int}}$ to perform explicit intermediate reasoning—analyzing trade-offs and resolving conflicts (e.g., adopting the logical predicate from one attempt while utilizing the contextual framing of another)—before generating the final output y^*:

$$y^* = \text{argmax}_y P_{\mathcal{M}_{\text{int}}}(y \mid \mathcal{C}, \mathcal{X}_{\text{ref}}, \mathcal{I}_{\text{fuse}}) \tag{4}$$

This process ensures the output is a logical composition rather than simple selection, effectively synthesizing a solution that surpasses any individual candidate in the pool.

3.4 Phase III: Distilled Integration

Although the proposed Explore-then-Integrate framework ensures high-quality synthesis, the dependency on a computationally intensive integrator creates significant latency bottlenecks. To circumvent this, we propose Distilled Integration, aiming to compress the integration logic—specifically the capability to arbitrate conflicts and synthesize fragments—into a compact student model $\mathcal{M}_{\text{stu}}$ (e.g., Llama-3-8B (Dubey et al., 2024)). We construct a training corpus $\mathcal{D}_{\text{sft}} = \{(\mathcal{X}_{\text{ref}}^{(i)}, y^{*(i)})\}$ pairing the structured reference contexts with the integrator's synthesized outputs. Distinct from conventional distillation pipelines that filter teacher outputs via external heuristics (Hsieh et al., 2023; Magister et al., 2023)—a process that risks re-imposing homogeneity bias—we adopt an unfiltered distillation strategy. By training on the full distribution of the integrator's decisions, we ensure $\mathcal{M}_{\text{stu}}$ internalizes the robust integration policy required to resolve complex constraints rather than merely memorizing instances. We fine-tune $\mathcal{M}_{\text{stu}}$ to minimize the standard negative log-likelihood (NLL) objective.

3.5 Provenance-aware Evaluation Suite

To validate that our method performs true compositional integration rather than trivial selection, and to confirm that high-entropy exploration exposes resolvable conflicts rather than introducing uncontrollable noise, we propose a fine-grained evaluation suite based on semantic fragment provenance. We decompose the output y and the candidate pool $\mathcal{H}_{\text{top}}$ into atomic semantic fragments (e.g., reasoning steps) and establish a provenance mapping via embedding cosine similarity. This granular attribution supports three metrics (formal definitions in Appendix B):

Source Entropy quantifies the diversity of information use. High entropy indicates effective synthesis of insights from multiple explorers rather than a collapse into a single dominant chain.

Multi-Source Interleaving Rate (MSIR) measures the granularity of integration. A higher MSIR indicates a more complex "jigsaw-style" composition process in which the model actively alternates between sources, rather than block-level copying.

Conflict Suppression Rate (Δ_{supp}) assesses the rationality of integration. We first calculate the *Conflict Density* $\delta(u)$ for each fragment u in $\mathcal{H}_{\text{top}}$ using a Subject-Verb-Object (SVO) proxy to identify contradictions with other chains. Δ_{supp} then measures the relative reduction in conflict density of the selected fragments compared with the discarded pool (i.e., $(\delta_{\text{dis}} - \delta_{\text{sel}})/\delta_{\text{dis}}$). A positive score indicates that the integrator acts as a semantic filter, effectively pruning logical noise.

4 Experiments

4.1 Settings

Datasets and Metrics. We evaluate our framework on CommonGen-Lite (Lin et al., 2020) and DimonGen (Liu et al., 2023), which require composing discrete concepts into coherent everyday scenarios. Adhering to established protocols, we report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and SPICE (Anderson et al., 2016) to assess coverage, rationality, and quality. Additionally, we incorporate NoRa (Zhou et al., 2024) to assess robustness to interference; this benchmark challenges models to maintain reasoning stability in the presence of noisy rationales containing irrelevant or inaccurate steps. For NoRa, we report standard Accuracy to quantify the model's ability to filter distractors and draw correct answers. Refer to Appendix C for dataset details.

Implementation Details. Guided by our preliminary analysis in Figure 2, which shows that integration efficacy correlates strongly with deep semantic diversity rather than surface-level lexical variation, we prioritize high-semantic-entropy sources in the exploration phase. Accordingly, we use Qwen-3-32B-Instruct (Bai et al., 2023) as the Explorer, paired with GPT-4o as the Integrator for compositional synthesis. FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023; Xiong et al., 2024) serves as the PRM for filtering, and LLaMA-3.1-8B-Instruct (Dubey et al., 2024) is used as the base model for the distilled student. We adopt an asymmetric sampling budget: the Explorer generates $M = 50$ candidates at $T = 0.6$ (filtered to the top-$K = 5$ via PRM), while the Integrator performs a single synthesis step. We compare our heterogeneous configuration (Hetero) against homogeneous variants (Homo, which use the same model for both roles). Additional baselines include standard few-shot CoT, a single-chain reasoning method (Direct), and the task-specific SOTA CD-CoT for the NoRa benchmark. Due to the space limitation, we put the detailed descriptions of baselines and variants in Appendix D.

4.2 Results

In this section, we provide a comprehensive evaluation of the Explore-then-Integrate framework. We structure our analysis around the following research questions (RQs):

RQ1. Can our framework achieve superior performance across diverse benchmarks, covering both standard GCR tasks and robustness-critical scenarios involving noisy rationales?

RQ2. Can the distilled student successfully internalize a generalizable integration policy, rather than merely memorizing specific teacher solutions?

RQ3. How does the choice of models for Exploration and Integration affect the framework's efficacy?

4.2.1 RQ1: Overall Performance and Robustness

Table 1 presents results on CommonGen-Lite. In the homogeneous configurations (Block I), the data corroborate the saturation and degradation trends characteristic of the homogeneity trap; notably, filtering a large homogeneous pool ($M = 50, K = 5$) reduces quality even below the single-chain baseline. Conversely, our heterogeneous configuration (Block II) yields substantial improvements across all metrics, simultaneously optimizing lexical coverage (BLEU-4, ROUGE-L) and semantic coherence (METEOR, SPICE), as evidenced by substantial margins of +7.25 BLEU-4 and +4.06 SPICE over the direct baseline.

As shown in Table 2, our heterogeneous configuration demonstrates superior robustness on NoRa, achieving a SOTA overall accuracy of 90.20% and outperforming the task-specific CD-CoT baseline by a wide margin. Crucially, under inaccurate noise—where rationales contain plausible but logically false premises—our method maintains

Method	B-4	SPICE	R-L	MET
Baseline: GPT-4o Direct	21.21	38.41	56.47	52.84
Block I. Homogeneous Configuration				
Exp-Int (Homo, GPT-4o, $M = 5$)	23.87	40.71	53.52	59.24
Exp-Int (Homo, GPT-4o, $M = 20$)	21.69	39.25	52.32	59.80
Exp-Int (Homo, GPT-4o, $M = 30$)	22.97	39.23	52.11	60.31
Exp-Int (Homo, GPT-4o, $M = 40$)	22.85	39.35	52.65	60.12
Exp-Int (Homo, GPT-4o, $M = 50$)	22.75	39.29	52.97	59.62
Exp-Int (Homo, GPT-4o, $M = 50, K = 5$)	19.43	37.81	50.51	58.26
Block II. Heterogeneous Configuration (Ours)				
Exp-Int (Hetero, Qwen3-32B + GPT-4o)	28.46	**42.47**	58.38	**61.94**
Exp-Int (Qwen3-32B + Dstl. LLaMA3-8B)	27.98	42.40	57.77	60.99
Block III. Integrator Ablation				
Exp-Int (Hetero, Qwen3-32B + DPSK-v3)	**29.03**	41.99	**59.87**	60.35
Exp-Int (Qwen3-32B + Dstl. Qwen3-8B)	27.51	41.70	56.72	60.44
Block IV. Explorer Ablation				
Exp-Int (Hetero, Qwen3-8B + GPT-4o)	24.96	41.26	54.90	59.78
Exp-Int (Hetero, LLaMA3-8B + GPT-4o)	24.68	40.70	54.84	59.71
Exp-Int (Hetero, LLaMA3-70B + GPT-4o)	22.53	39.39	52.78	57.58

Table 1: Performance comparison on **CommonGen-Lite**. We compare Homogeneous configurations (Block I) with our proposed Heterogeneous framework (Block II), and present ablation studies on the Integrator (Block III) and Explorer (Block IV). Abbreviations: B-4 (BLEU-4), R-L (ROUGE-L), MET (METEOR), Exp-Int (Explore-then-Integrate), and Dstl. (Distilled). The best scores are highlighted in **bold**, and the second-best scores are underlined.

89.82% accuracy. We attribute this robustness to the synergy of broad exploration and granular fusion: the high-semantic-entropy explorer searches widely for valid reasoning paths to override local noise, while the integrator performs fragment-level synthesis. This fine-grained composition enables the model to precisely identify and prune logical incoherence, effectively decoupling valid reasoning from contaminated context.

Since lexical metrics alone may not fully reflect semantic preference, we further conduct an LLM-as-a-Judge evaluation on CommonGen-Lite. We follow a double-blind pairwise protocol and additionally report sentence length, concept coverage, and part-of-speech correctness. As shown in Table 3, the heterogeneous configuration achieves a substantially higher judge win rate than the homogeneous baseline, while also yielding shorter generations, better concept coverage, and higher part-of-speech accuracy. These results indicate that the gains of our framework are not limited to lexical overlap, but are also reflected in semantic preference and logical quality.

Notably, this superiority generalizes across domain boundaries. As detailed in Appendix E (Table 7), our framework achieves best-in-class performance across all tasks, including Math, Symbolic, and Commonsense. Even on the challenging Symbolic-Longer task, where the baseline collapses to 62.00%, our method recovers to 78.00%. This cross-domain consistency underscores the potential of Explore-then-Integrate not merely as a GCR solution, but as a universally robust reasoning paradigm.

4.2.2 RQ2: Effectiveness of Distilled Integration.

Table 1 (Block II) shows that the distilled student model closely matches its teacher's performance. Despite a drastic reduction in parameter count (from GPT-4o to 8B), the student achieves 42.40 SPICE and 27.98 BLEU-4, with negligible degradation relative to the complete heterogeneous system (42.47 SPICE, 28.46 BLEU-4). This proves that the complex integration logic can be effectively compressed into a compact local model without sacrificing reasoning fidelity.

To determine whether the student merely memorizes specific solution patterns or acquires a generalizable fusion policy, we evaluate its zero-shot performance on DimonGen using the LoRA weights trained solely on CommonGen-Lite. As shown in Table 4, the student exhibits remarkable transfer capabilities. It surpasses the GPT-4o Direct baseline on BLEU-4 and ROUGE-L. This performance, achieved without any target-domain fine-tuning, indicates that the student has successfully internalized an abstract integration logic.

4.2.3 RQ3: Integrator Generalization

Results in Table 1 (Block III) assess the framework's adaptability when replacing GPT-4o with DeepSeek-v3 (DPSK-v3) (DeepSeek-AI, 2024) as the integrator. While DPSK-v3 scales better with homogeneous sampling than GPT-4o, introducing heterogeneity via the Qwen3-32B explorer still yields the highest BLEU-4 and ROUGE-L. Furthermore, we validate the architecture-agnostic nature of our distillation pipeline: a student model based on Qwen3-8B achieves competitive performance, mirroring the success of the LLaMA3-8B student. This demonstrates that the student's ability to internalize the fusion policy is robust to the underlying model architecture.

4.2.4 RQ3: The Alignment-Entropy Trade-off of Explorer

Building on our finding in Figure 2 that deep semantic diversity drives final performance, Table

Method	overall Acc. (%)	by difficulty Easy	by difficulty Hard	by noise type Inacc.	by noise type Irrel.	an example task: Symbolic-L
Baseline: GPT-4o Direct	82.67	81.60	82.80	81.68	83.75	62.00
Baseline: CD-CoT	80.00	80.80	80.60	81.42	78.43	51.33
Exp-Int (Homo, GPT-4o)	85.60	88.40	83.80	84.86	86.41	66.00
Exp-Int (Hetero, Qwen3-32B + GPT-4o)	**90.20**	**90.60**	**89.60**	**89.82**	**90.62**	**78.00**

Table 2: Accuracy comparison on **NoRa**. Results are categorized by (1) difficulty (determined by noise ratios), (2) noise type (Inaccurate/Irrelevant rationales), and (3) task type, represented here by the challenging Symbolic-Longer task. Full breakdowns across all tasks are provided in Table 7.

Configuration	Length	Coverage	PoS	Judge Win Rate
Exp-Int (Homo)	13.16	97.00%	92.25%	28.6%
Exp-Int (Hetero)	**12.11**	**98.25%**	**95.00%**	**59.0%**

Table 3: LLM-as-a-Judge and logical quality comparison on **CommonGen-Lite**. We conduct double-blind pairwise evaluation with swapped presentation order for each sample pair. Length denotes the average number of words. Coverage measures the proportion of generations that include all concepts, and PoS measures the proportion with correct part-of-speech usage. The remaining 12.4% of comparisons are ties.

Method	B-4	SPICE	R-L	MET
Baseline: GPT-4o Direct	7.93	**14.89**	32.29	**35.81**
Exp-Int (Homo, GPT-4o)	7.07	13.75	30.22	34.07
Exp-Int (Qwen3-32B + Dstl. LLaMA3-8B)	**10.12**	14.43	**33.97**	34.26

Table 4: Cross-dataset generalization results on **DimonGen**. The student model (Dstl. LLaMA3-8B) is trained exclusively on CommonGen-Lite and evaluated zero-shot on DimonGen to assess the transferability of the learned fusion policy.

1 (Block IV) provides a concrete perspective on how explorer characteristics influence this property. We observe that the highly capable LLaMA3-70B significantly underperforms its smaller 8B counterpart and the Qwen3-32B model (e.g., 22.53 vs. 24.96 and 28.46 BLEU-4). This supports an inverse scaling law for exploration: as models scale up and undergo rigorous alignment, their utility as explorers often declines. Such alignment tends to compress the probability distribution, effectively pruning the distributional tail necessary for capturing novel concept bindings (Lin et al., 2024; Huang et al., 2025). Consequently, this confirms that the decisive attribute for an explorer is not raw reasoning power but the capacity to maintain an uninhibited semantic search space that provides complementary reasoning paths to the integrator.

We further study how the sampling budget affects the final performance of the heterogeneous explorer. Specifically, we vary the number of sampled candidates M and track both diversity and downstream generation quality. As shown in Table 5, increasing M from 5 to 20 steadily improves both entropy and semantic quality, with SPICE reaching its peak at $M = 20$. Beyond this point, the gains become marginal and the semantic metric begins to fluctuate, although BLEU-4 still shows a mild increase. This suggests that a moderate sampling budget already provides sufficient semantic coverage after PRM filtering, and we therefore adopt $M = 20$ as a practical default in the main experiments.

M	Entropy (pre-PRM)	Entropy (post-PRM)	B-4	SPICE
5	0.93	0.93	25.54	41.98
10	1.19	0.93	26.73	42.32
15	1.33	0.94	26.84	42.55
20	**1.41**	**0.96**	27.30	**42.67**
30	1.53	0.98	27.36	41.69
50	1.66	0.99	**28.46**	42.47

Table 5: Effect of the sampling budget M on diversity and performance. We vary the number of sampled candidates in the exploration phase and report diversity before and after PRM filtering, together with the final performance on **CommonGen-Lite**.

5 Further Analysis

Although the preceding experiments establish the efficacy of the Explore-then-Integrate framework, strictly validating the source of these gains requires examining the underlying mechanisms. In this section, we move beyond performance metrics to address two fundamental questions: (1) How does diversity drive the utility of exploration? (2) What underlying mechanisms govern the "black box" of the integration?

5.1 Unveiling the Explorer Potential

To elucidate the mechanisms behind the homogeneity trap, we trace the evolution of search space diversity from the initial candidate pool ($\mathcal{H}$) to the filtered subset ($\mathcal{H}_{\text{top}}$). We present the results in Appendix G (Table 9) using comprehensive metrics to distinguish between trivial lexical variation and substantive semantic diversity.

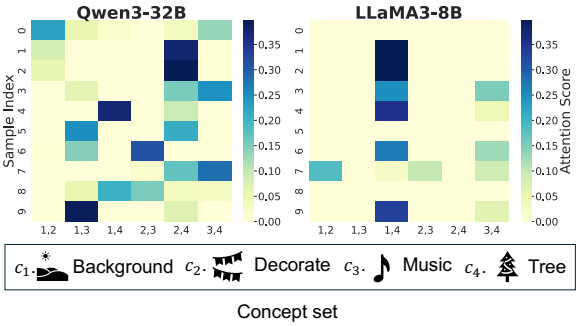

Figure 4: Heatmaps comparing pairwise concept attention between Qwen3-32B and LLaMA-8B across 10 random samples. Qwen3-32B covers the full spectrum of potential concept bindings. LLaMA3-8B narrowly fixates on specific relations (e.g., c_1-c_4) while completely neglecting valid alternatives (e.g., c_1-c_3, c_2-c_4).

Surface-level Variation. We examine lexical diversity using Entropy-4 (Li et al., 2016) to quantify the uniformity of 4-gram distributions and Self-BLEU4 (Zhu et al., 2018) to measure text overlap. In $\mathcal{H}$, smaller models such as LLaMA-3-8B exhibit high entropy and low overlap, in contrast to the rigid, template-like generation of aligned models such as DPSK-v3. However, the performance gap observed in our main experiments suggests that this surface variation is largely orthogonal to final performance. Merely rephrasing the same logic provides trivial benefits to the integrator, serving as syntactic noise rather than information gain.

Deep Semantic Diversity. The key distinction emerges in the semantic space, probed via three metrics. We employ Self-CosSim (Cox et al., 2021) to quantify content-level overlap using the average pairwise embedding similarity. We use the Vendi Score (VS) (Friedman and Dieng, 2023; Pasarkar and Dieng, 2024), derived from the eigenvalues of the similarity kernel matrix, and report orders $q = \{0.5, 1, \infty\}$ to assess sensitivity across rare tails and dominant modes. Additionally, we compute Embedding Entropy from the RBF Gram spectrum (Friedman and Dieng, 2023). All semantic metrics use embeddings obtained via Sentence-BERT (Reimers and Gurevych, 2019).

In $\mathcal{H}$, Qwen3-32B defines a vastly superior search space, achieving the lowest Self-CosSim and the highest VS, effectively occupying the distributional tail compared with the narrow semantic band of GPT-4o. In $\mathcal{H}_{\text{top}}$, diversity scores show consistency, aligning closely across all models (e.g., Self-CosSim ≈ 0.88). Rather than negating our hypothesis, this reflects the PRM's necessary role in pruning irrational hallucinations and guiding the pool toward valid commonsense truth. The critical distinction lies in the nature of this consistency. For homogeneous models, low diversity arises from premature convergence: within a restricted bias band, selected candidates are merely redundant variants of a local mode. Conversely, our framework achieves robust convergence: the PRM acts on a broad initial search to identify the global optimum. Thus, the Explorer's value lies not in enforcing diversity in the final output, but in ensuring that the initial search space is sufficiently expansive to prevent the Integrator from becoming trapped in local homogeneity.

To visually distinguish effective semantic exploration from deceptive lexical variation, Figure 4 compares the pairwise concept attention heatmaps of Qwen3-32B and LLaMA3-8B. Qwen3-32B exhibits a dispersed distribution with high variance, indicating broad exploration of diverse concept bindings. However, LLaMA3-8B shows a repetitive focus on the same bindings despite its high surface-level diversity, suggesting that its variation is confined to syntax rather than semantics.

5.2 Unpacking the Integration Process

To investigate the nature of integration, we propose the **Provenance-aware Evaluation Suite**. By tracing the semantic origins of generated fragments, we quantify three dimensions: Source Entropy (diversity of utilization), MSIR (granularity of fusion), and Conflict Suppression (Δ_{supp}), which measures the relative reduction in logical contradictions between selected and discarded fragments.

As shown in Table 6, the homogeneous configuration yields a low MSIR (32.2%) and Source Entropy (0.36), suggesting that when inputs share systematic biases, the integrator defaults to block-level copying. In contrast, the heterogeneous framework significantly elevates Source Entropy to 0.58 and MSIR to 50.5%. Rather than struggling to reconstruct missing logic from scratch, the integrator acts as a "jigsaw solver", identifying and weaving together complementary fragments to construct a narrative superior to any single source.

Crucially, our results address the concern that high-entropy explorers might introduce uncontrollable reasoning noise. Empirically, the heterogeneous framework achieves a significantly lower Conflict Density (Selected) than the homogeneous setup (0.38 vs. 0.49). This shows that potential conflicts within a heterogeneous pool are *explicitly*

Method	Source Entropy ↑	MSIR (%) ↑	Conflict Density (Selected) ↓	Conflict Density (Discarded)	Δ_{supp} (%) ↑
Exp-Int (Homo)	0.36	32.2	0.49	0.58	15.52
Exp-Int (Hetero)	**0.58**	**50.5**	0.38	0.45	15.56
Exp-Int (w/ Dstl.)	0.57	49.5	**0.37**	0.45	**17.78**

Table 6: Provenance-aware evaluation results. We quantify the diversity of utilization (Source Entropy), the granularity of integration (MSIR), and the efficacy of conflict suppression (Δ_{supp}) across homogeneous, heterogeneous, and distilled configurations.

exposed, making them easier to extract, compare, and arbitrate. However, homogeneous chains often fail to benefit from their individual quality; instead, they mask core logical flaws behind a veneer of consensus, thereby increasing the integrator's burden of identifying hidden contradictions. Consequently, introducing high-semantic-entropy explorers does not introduce unmanageable noise but rather facilitates the precise pruning of irrationality.

Finally, the provenance metrics also provide a mechanistic explanation for the efficacy of the Distilled Student, distinguishing its behavior from simply memorizing surface patterns.

6 Conclusion

In this work, we uncover the "Homogeneity Trap" in GCR, where scaling homogeneous ensembles degrades performance due to semantic mode collapse. Our mechanistic investigation reveals that deep semantic diversity—rather than surface-level lexical variation—is the decisive prerequisite for effective integration. To escape this trap, we propose the Explore-then-Integrate framework, which synergizes a high-semantic-entropy explorer with a powerful integrator. By deploying our Provenance-aware Evaluation Suite, we rigorously verify that our performance gains stem from granular logical composition and active conflict suppression. Extensive experiments demonstrate SOTA results across diverse benchmarks (e.g., 90.2% accuracy on NoRa). Furthermore, our Distilled Integration shows that this complex integration policy can be effectively internalized into a compact student model, enabling efficient, standalone inference. Future work includes extending this paradigm to iterative and long-form reasoning scenarios, such as self-refine and exchange-of-thoughts (Madaan et al., 2023; Yin et al., 2023).

Limitations

Our approach performs compositional synthesis in a single turn. Currently, the integrator merges fragments to satisfy constraints in one pass. However, for complex or long-form reasoning tasks, a single step may not suffice to resolve all logical conflicts. Extending this paradigm to iterative settings, such as self-refinement or multi-turn interactions, remains to be explored to further enhance reasoning stability.

7 Acknowledgments

This work was supported in part by Natural Science Foundation of China (No. U24B6012, 61972217, 32071459, 62176249, 62006133, 62271465, 62406167), the Shenzhen Medical Research Funds in China (No. B2302037), the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122702), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Ming Li, Likang Xiao, Dingqi Yang, and 1 others. 2025. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*.

Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human*

Factors in Computing Systems, CHI '21, New York, NY, USA. Association for Computing Machinery.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. MORE: Multi-mOdal REtrieval augmented generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1178–1192, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. Preprint.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kulkarni, Tiago Ramalho, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, page 93–108.

Neel Guha, Mayee Chen, Trevor Chow, Ishan Khare, and Christopher Re. 2024. Smoothie: Label free language model routing. *Advances in Neural Information Processing Systems*, 37:127645–127672.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. John Benjamins Philadelphia, PA.

Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Sm Yiu, and Nan Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hsiang-Fu Nakayama, James Lee, and Ankur Taly. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5403–5420.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. KFCNet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2918–2928, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need. *arXiv preprint arXiv:2402.05120*.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When thinking fails: The pitfalls of reasoning for instruction-following in LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang

Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023. DimonGen: Diversified generative commonsense reasoning for explaining concept relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4719–4731, Toronto, Canada. Association for Computational Linguistics.

Xin Liu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Junwei Ding, Wenqing Yao, Weihua Luo, Haiying Zhang, and Jinsong Su. 2022. Kgr4: Retrieval, retrospect, refine and rethink for commonsense generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11029–11037.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. KG-BART: Knowledge graph-augmented BART for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8411–8419.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*.

Bo Lv, Chen Tang, Yanan Zhang, Xin Liu, Ping Luo, and Yue Yu. 2024. URG: A unified ranking and generation method for ensembling language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4421–4434, Bangkok, Thailand. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1773–1796. Also arXiv:2212.08410 (2022).

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amey P Pasarkar and Adji Bousso Dieng. 2024. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3808–3816. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Jacob Russin, Sam Whitman McGrath, and Danielle J. Williams. 2025. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *Preprint*, arXiv:2405.15164.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A survey on compositional learning of AI models: Theoretical and experimental practices. *Transactions on Machine Learning Research*. Survey Certification.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc 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.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2025. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *Preprint*, arXiv:2312.11456.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2025. Evaluating the evaluation of diversity in commonsense generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24258–24275, Vienna, Austria. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023a. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Denny Zhou, Dale Schuurmans, Xuezhi Wang, Jason Huang, Jason Wei, Ed Chi, and Quoc Le. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Detailed Instructions

A.1 Explorer Instruction

[Explorer Instruction] :

[Task Definition] : You must start your response with `<think>` and then think step by step. Given several concepts (nouns/verbs), write a short, simple sentence describing a common daily scene containing all required words.

[Few-shot Demonstrations] :
Example 1: Concepts: "dog, frisbee, catch, throw" → Sentence: The dog catches the frisbee...
Example 2: Concepts: "apple, place, tree, pick" → Sentence: A girl picks some apples...

[Input] : Concepts: `{concepts}`

A.2 PRM Instruction

[PRM Scoring Criteria] :

[Role Definition] : As an AI output quality assessment expert, please comprehensively evaluate the following dimensions.

[Evaluation Metrics] :

1. **Language Coherence**: Sentence structure, natural expression, logical clarity.
2. **Length Appropriateness**: Reasonable length and reasoning chain.
3. **Keyword Coverage**: Whether all `{keywords}` are used accurately.
4. **Integration Density**: Natural incorporation avoiding forced assembly.

[Output Format] : Provide a comprehensive score based on these criteria.

A.3 Integrator Instruction

[Context Injection] :

[System Instruction] : The following are previous attempts generated for the keywords `{keywords}`.

[Reference Attempts] :
Attempt 1: `{output_from_chain_1}`
Attempt 2: `{output_from_chain_2}`
...

[Fusion Instruction] :

[Goal] : Create the best possible output combining useful information from the reference attempts.

[Fusion Strategy] :

- Combine useful word combinations and sentence structures from different CoT attempts.
- Extract the best patterns and approaches from the references.
- Aim for natural, grammatically correct expression.

[Constraints] :

- Use **ONLY** the given keywords: `{keywords}`.
- Start with `<think>` to show your conflict resolution process.
- Provide final answer in `<answer>` tags.

B Detailed Provenance-aware Evaluation Suite

We first decompose the final generation y into a sequence of logical fragments $(y_1, \ldots, y_T)$ and similarly flatten the filtered candidate pool $\mathcal{H}_{\text{top}}$—comprising K chains—into a set of source fragments $\mathcal{U}_{pool} = \{h_{k,j}\}$ (denoting the j-th fragment of the k-th chain) using a syntactic dependency parser, spaCy. Let $\mathbf{e}(\cdot)$ denote the embedding vector derived from a sentence transformer (Reimers and Gurevych, 2019). We define the provenance mapping $\phi(y_t)$ for a generated fragment y_t as the index of the source chain containing the best-matching fragment, subject to a validity threshold τ:

$$\phi(y_t) = \begin{cases} \arg\max_k \left(\max_j \cos(\mathbf{e}(y_t), \mathbf{e}(h_{k,j})) \right) \\ \qquad \text{if } \max_{k,j} \cos(\cdot) > \tau \\ \emptyset \qquad \text{otherwise} \end{cases}$$

Based on this mapping, we introduce three metrics. First, to quantify the diversity of information utilization, we compute the **Source Entropy**. Let $n_k = \sum_{t=1}^{T} \mathbb{I}(\phi(y_t) = k)$ be the count of fragments attributed to chain k. The metric is defined as the normalized Shannon entropy: $-\frac{1}{\log K} \sum_{k=1}^{K} \frac{n_k}{T'} \log \frac{n_k}{T'}$, where T' is the total number of attributed fragments. High entropy implies the integrator effectively synthesizes insights from multiple explorers. Second, to capture the granularity of fusion, we measure the **Multi-Source Interleaving Rate** (MSIR), which measures the frequency of source transitions between adjacent attributed fragments:

$$\text{MSIR} = \frac{\sum_{t=1}^{T-1} \mathbb{I}\begin{pmatrix} \phi(y_t) \neq \emptyset \wedge \phi(y_{t+1}) \neq \emptyset \\ \wedge \ \phi(y_t) \neq \phi(y_{t+1}) \end{pmatrix}}{\sum_{t=1}^{T-1} \mathbb{I}\big(\phi(y_t) \neq \emptyset \wedge \phi(y_{t+1}) \neq \emptyset \quad \big)}.$$

A higher MSIR indicates a complex jigsaw-style composition process rather than block-level copying. Finally, to verify the rationality of integration, we calculate the **Conflict Suppression Score** (Δ_{supp}). We utilize a lightweight proxy based on Subject-Verb-Object (SVO) extraction to measure logical inconsistency. For any source fragment $u \in \mathcal{U}_{pool}$ belonging to a chain, we define its Conflict Density $\delta(u)$ as the proportion of fragments in other chains that share entity overlaps but contain

contradictions:

$$\delta(u) = \frac{\sum_{\substack{v \in \mathcal{U}_{pool} \\ k(v) \neq k(u)}} \mathbb{I}\begin{pmatrix}\text{Overlap}(u,v) \\ \wedge\ \text{Contradicts}(u,v)\end{pmatrix}}{\sum_{\substack{v \in \mathcal{U}_{pool} \\ k(v) \neq k(u)}} \mathbb{I}(\text{Overlap}(u,v))},$$

where $k(\cdot)$ extracts the chain index. Here, Overlap requires identical subject or object lemmas, while Contradicts checks for divergent verbs or negation polarity. We then partition the pool into two subsets: Selected Set $\mathcal{F}_{sel} = \{u \in \mathcal{U}_{pool} \mid \exists t, \phi(y_t) = u\}$ containing fragments inherited by the final output, and the Discarded Set $\mathcal{F}_{dis} = \mathcal{U}_{pool} \setminus \mathcal{F}_{sel}$. The suppression score quantifies the relative reduction in conflict density of the retained fragments compared to the discarded ones:

$$\Delta_{supp} = \frac{\mathbb{E}_{u \in \mathcal{F}_{dis}}[\delta(u)] - \mathbb{E}_{u \in \mathcal{F}_{sel}}[\delta(u)]}{\mathbb{E}_{u \in \mathcal{F}_{dis}}[\delta(u)]}.$$

A positive Δ_{supp} indicates that the integration policy effectively suppresses high-conflict fragments, retaining a set whose semantic inconsistency is proportionally lower than that of the discarded pool.

C Dataset Details

We evaluate our framework on three challenging benchmarks, each targeting a distinct aspect of reasoning:

- CommonGen-Lite (Hard) (Lin et al., 2020) is a widely adopted benchmark for GCR. The task requires models to generate a coherent sentence describing an everyday scenario using a given set of input concepts (e.g., {dog, frisbee, catch, throw}). We use the *Hard* split, in which concept sets have low co-occurrence frequencies.
- DimonGen (Liu et al., 2023) extends GCR by focusing on the *diversity* of concept relationships. It requires generating multiple distinct narratives for a concept pair (e.g., "dog" and "sheep") that depict unique relational perspectives (e.g., herding vs. attacking).
- NoRa (Zhou et al., 2024) evaluates reasoning resilience against Noisy Rationales, where context examples contain *irrelevant* (extraneous facts) or *inaccurate* (logical fallacies) reasoning steps across mathematical, symbolic, and commonsense domains.

D Baseline and Variant Descriptions

To strictly isolate the effects of source diversity, candidate scaling, and model architecture, we categorize our experimental comparisons into baselines and four blocks of framework instantiations.

Baselines. We report the performance of standard models and task-specific methods to establish competitive bounds:

- GPT-4o Direct: The standard single-chain reasoning baseline with few-shot CoT prompting.
- CD-CoT (Contrastive Denoising with Noisy CoT) (Zhou et al., 2024): The state-of-the-art method specifically designed for the NoRa benchmark. CD-CoT operates on the premise that models can rectify rationales by contrasting noisy examples with clean ones. It executes a complex four-step pipeline: rationale rephrasing, rationale selecting, rationale exploring, and answer voting. The first two steps aim to achieve explicit denoising, while the last two steps are for diverse reasoning paths. This process is computationally intensive, typically requiring approximately 14 GPT-4o API calls per instance.

Framework Instantiations. All variants adhere to the proposed Explore-then-Integrate framework: generating M candidates via an Explorer, filtering to K via a PRM (unless specified otherwise), and synthesizing via an Integrator. We categorize the configurations into four blocks corresponding to Table 1:

- Block I: Homogeneous Configuration. GPT-4o serves as both the Explorer and the Integrator. To analyze scaling effects, we evaluate: (1) *Standard* ($M = 5$); (2) *Scaled-Full* ($M = 20, 30, 40, 50$, all passed to integrator); and (3) *Scaled-Filtered* ($M = 50$, filtered to $K = 5$).
- Block II: Heterogeneous Configuration. Our primary setup pairs Qwen3-32B (Explorer) with GPT-4o (Integrator). This block also includes the Distilled Student, where LLaMA3-8B is fine-tuned to emulate the primary heterogeneous teacher.
- Block III: Integrator Ablation. To test model agnosticism, we replace GPT-4o with DeepSeek-v3 and use an alternative Distilled Student based on Qwen3-8B.

- Block IV: Explorer Ablation. Fixing GPT-4o as the Integrator, we vary the Explorer architecture to verify the "Alignment-Entropy Trade-off". Models include Qwen3-8B, LLaMA3-8B, and the heavily aligned LLaMA3-70B.

E Detailed Results on NoRa

Table 7 presents the performance breakdown across all five tasks in the NoRa benchmark. These results confirm that our framework achieves robust generalization across mathematical, symbolic, and commonsense domains.

F Analysis of Hybrid Explorer Ensembles

To investigate the impact of mixing explorers with different characteristics, we constructed hybrid ensembles by combining candidates from multiple models. Keeping a fixed integrator budget of $M = 5$, we retained the top-3 candidates from our primary explorer (Qwen3-32B) and supplemented them with the top-2 candidates from an auxiliary model (LLaMA3-8B or LLaMA3-70B).

As shown in Table 8, this hybrid strategy consistently underperforms the pure Qwen3-32B baseline, suggesting that simply diversifying model sources does not guarantee improvement:

- Mixing LLaMA3-8B: Performance drops to 25.60 BLEU-4. Although LLaMA3-8B is a high-entropy model, its top candidates do not match the semantic utility of the Qwen3-32B candidates they replace (i.e., Qwen's rank 4 and 5), thereby degrading the overall quality of the candidate pool.

- Mixing LLaMA3-70B: Performance degrades further to 20.86. Consistent with our findings in the main text regarding the Alignment-Entropy Trade-off, introducing the heavily aligned 70B model reduces the overall semantic entropy of the context window, limiting the diversity necessary for effective fusion.

G Analysis of Diversity Metrics

Table 9 provides the detailed quantitative breakdown of the search space evolution discussed in Section 5, contrasting surface-level diversity against deep semantic diversity across different explorer architectures.

Method	Math-Base9	Math-Base11	Symbolic-Equal	Symbolic-Longer	Commonsense	Overall
Baseline: GPT-4o Direct	95.33	88.00	87.00	62.00	81.00	82.67
Baseline: CD-CoT	98.67	92.33	75.33	51.33	82.33	80.00
Exp-Int (Homo, GPT-4o)	98.67	93.00	**88.33**	66.00	82.00	85.60
Exp-Int (Hetero, Qwen3-8B + GPT-4o)	**100.00**	**98.67**	87.00	**78.00**	**87.33**	**90.20**

Table 7: Performance breakdown on NoRa across different tasks. Our heterogeneous framework achieves the best or second-best performance on all tasks, with significant margins on the hardest task (Symbolic-Longer), demonstrating universal robustness.

Pool Composition ($M = 5$)	**B-4**	**SPICE**	**R-L**	**MET**
Pure Ensemble				
Qwen3-32B (Top-5)	**28.46**	**42.47**	**58.38**	**61.94**
Hybrid Ensembles				
Qwen3-32B (Top-3) + LLaMA3-8B (Top-2)	25.60	41.55	55.25	60.92
Qwen3-32B (Top-3) + LLaMA3-70B (Top-2)	20.86	38.90	51.13	58.57

Table 8: Performance comparison on Commongen-Lite between the pure Qwen-32B explorer and hybrid ensembles mixed with LLaMA models.

Model	**Entropy-4**↑	**Self-BLEU4**↓	**Self-CosSim**↓	**VS-Embed**$_{0.5}$↑	**VS-Embed**$_{1}$↑	**VS-Embed**$_{\infty}$↑	**Embed-Entropy**↑
Candidate Pool Diversity ($\mathcal{H}, M = 50$)							
GPT-4o	10.69	79.15	0.87	0.13	0.13	0.32	1.72
DPSK-v3	9.37	94.36	0.95	0.05	0.06	0.14	1.63
LLaMA3-70B	10.98	82.22	0.93	0.07	0.07	0.24	1.79
Qwen3-32B	10.75	81.62	**0.66**	**0.32**	**0.34**	**0.89**	1.66
Qwen3-8B	11.25	89.50	0.74	0.21	0.26	0.75	1.52
LLaMA3-8B	**11.50**	**64.33**	0.87	0.13	0.13	0.48	**1.86**
Filtered Subset Diversity ($\mathcal{H}_{\text{top}}, K = 5$)							
GPT-4o	8.62	44.76	**0.87**	**0.12**	**0.13**	**0.19**	0.97
DPSK-v3	8.82	67.03	0.92	0.07	0.08	0.12	0.93
LLaMA3-70B	9.25	51.30	0.94	0.06	0.06	0.10	0.98
Qwen3-32B	9.97	49.92	0.88	**0.12**	0.12	0.18	**0.99**
Qwen3-8B	**10.03**	61.74	0.92	0.10	0.11	0.12	**0.99**
LLaMA3-8B	9.59	**42.88**	0.88	0.11	0.12	0.18	**0.99**

Table 9: Quantitative analysis of search space evolution on Commongen-Lite. We report surface-level (Entropy-4, Self-BLEU4) and deep semantic (Self-CosSim, VS-Embed, Embed-Entropy) diversity metrics for both the initial candidate pool ($\mathcal{H}, M = 50$) and the filtered subset ($\mathcal{H}_{\text{top}}, K = 5$).