LinkQA: Synthesizing Diverse QA from Multiple Seeds Strongly Linked by Knowledge Points

Xuemiao Zhang[1,3*], **Can Ren**[1,3*], **Chengying Tu**[2,3*],
Rongxiang Weng[3†], **Hongfei Yan**[2,4†], **Jingang Wang**[3], **Xunliang Cai**[3]
[1] Peking University [2] School of Computer Science, Peking University [3] Meituan
[4] State Key Laboratory of Multimedia Information Processing, Peking University
{zhangxuemiao, yanhf}@pku.edu.cn {tuchengying, 2401210098}@stu.pku.edu.cn
wengrongxiang@gmail.com {wangjingang02, caixunliang}@meituan.com

Abstract

The advancement of large language models (LLMs) struggles with the scarcity of high-quality, diverse training data. To address this limitation, we propose LinkSyn, a KP-graph-based synthesis framework that for the first time enables flexible control over discipline and difficulty distributions while balancing KP coverage and popularity. LinkSyn extracts KPs from question-answering (QA) seed data and constructs a KP graph to synthesize diverse QA data from multiple seeds strongly linked by KPs and sampled from graph walks. Specifically, LinkSyn incorporates (1) a knowledge value function to guide the adjustment of path sampling probability and balance KP coverage and popularity during graph walks; (2) diffusion-based synthesis via a strong reasoning model by leveraging multiple seeds with dense logical associations along each path; and (3) high-difficulty QA enhancement within given disciplines by flexible difficulty adjustments. By executing LinkSyn, we synthesize LinkQA, a diverse multi-disciplinary QA dataset with 50B tokens. Extensive experiments on Llama-3 8B demonstrate that continual pre-training with LinkQA yields an average improvement of 11.51% on MMLU and CMMLU, establishing new SOTA results. LinkQA consistently enhances performance across model size and initial FLOPs scales.[1]

1 Introduction

As the scale of LLMs escalates exponentially, the scarcity of high-quality training data has emerged as a critical bottleneck (Muennighoff et al., 2025; Villalobos et al., 2024), particularly in multi-disciplinary domains (Kandpal et al., 2023). Data synthesis has consequently gained prominence as a viable solution, offering scalable production of domain-specific knowledge representations (Gunasekar et al., 2023; Li et al., 2023; Nadăș et al., 2025). Crucially, synthetic data in QA format has demonstrated significant efficacy in enhancing model performance on knowledge-intensive tasks by providing structured reasoning pathways and explicit knowledge representations (Chen et al., 2025; Wang et al., 2025; Maini et al., 2024).

Despite these advances, current synthesis methods that depend on seed corpora face significant limitations. First, single-seed synthesis using trained models often experiences limited diversity due to inherent model biases (Su et al., 2025; Akter et al., 2025; Zhou et al., 2025). Second, entity-based methods (Qin et al., 2025; Jiang et al., 2025b; Yang et al., 2025), which extract sets of entities mentioned in documents and link documents by entity co-occurrence, aim to synthesize data from multiple connected documents. However, these methods exhibit limited knowledge integration, as individual entities rarely represent the document's core subject. Consequently, such connections lack semantic coherence, thus restricting cross-textual knowledge integration. Additionally, current methods struggle to finely adjust the distributions of synthesized data in terms of difficulty, discipline, and knowledge popularity. This results in a low yield of valuable data and poor performance on benchmarks that require higher-order abilities, such as reasoning (Hendrycks et al., 2021a).

Intuitively, unlike documents that mention numerous entities, a QA instance typically examines a few knowledge points (KPs), allowing each KP to serve as a strong representation of the QA itself. Thus, KP co-occurrence inherently provides more logical connections between QAs. Building on this insight, we construct a KP graph from QA seeds for the first time to capture robust logical associations. We then introduce LinkSyn, a novel diversity-driven and theoretically rigorous synthesis framework that traverses this graph to

[*] Equal contribution.
[†] Corresponding author.
[1] The core code and dataset are available at https://github.com/rc314159-creator/linkQA.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22620–22651
July 2-7, 2026

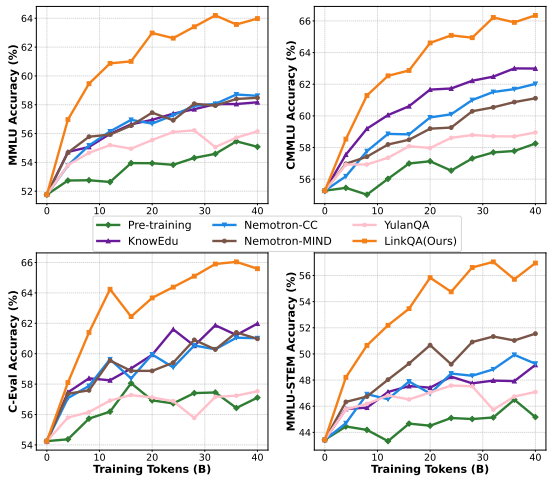

Figure 1: Comparison between LinkQA and baselines.

generate multi-seed QAs with dense logical links. Specifically, LinkSyn: (1) introduces a knowledge value-guided path sampler that balances structural alignment and long-tail exploration during graph walks; (2) synthesizes diverse or entirely novel QAs via a strong reasoning model in a diffusion-based manner by leveraging multiple seeds with dense logical associations along each path; and (3) enhances the concentration of high-difficulty QAs within specified disciplines during synthesis by flexibly adjusting the difficulty levels and discipline proportions of the sampling seed instances along graph paths.

We conduct extensive experiments by continually pre-training Llama-3 8B (Dubey et al., 2024) at the 2T-token checkpoint using 20B-token LinkQA, a multi-disciplinary QA dataset synthesized by LinkSyn. Our main contributions are summarized as follows: (1) For the first time, we construct a KP graph from KPs extracted from seed QA data, and introduce a theoretically rigorous framework, LinkSyn, to synthesize diverse multi-seed QAs with strong KP-based links. (2) We dedicate substantial resources to executing LinkSyn to synthesize LinkQA, a diverse multi-disciplinary QA dataset with 50B tokens. LinkQA is controllable in terms of difficulty, discipline, and KP distributions, fostering community research in scalable data synthesis and LLM advancement. (3) Extensive experiments conducted on Llama-3 8B, as shown in Figure 1, demonstrate that 20-B token LinkQA improves by an average of **11.51%** on MMLU and CMMLU over the pre-training baseline and achieves SOTA average performance across benchmarks, with scalable performance gains.

2 Method

2.1 Overview

LinkSyn is a framework designed to combine multiple seed QA instances to generate diverse samples while matching pre-specified *attribute* distributions (e.g., difficulty and discipline). The seeds are sourced through an extensive collection from open-source QA datasets. As illustrated in Figure 2, we first extract representative knowledge points (KPs) from each QA instance and construct a weighted KP graph, where edge weights quantify the strength of logical relations via KP co-occurrence statistics in the seed corpus. We then navigate the knowledge space using two complementary graph walking policies: a *popularity-priority* walk that samples neighbors proportional to edge weights (preserving local co-occurrence structure), and a *coverage-priority* walk that samples neighbors uniformly (encouraging exploration of long-tail connections). Conditioned on the visited KPs, we sample seed instances under the target difficulty and discipline constraints. Finally, our diffusion-based synthesis combines logically related instances along the sampled paths to create novel QA pairs while maintaining knowledge integrity.

2.2 Knowledge Point Graph Construction

Knowledge Point Extraction and Consolidation. We define KPs as the fundamental units of content within a discipline, such as concepts, principles, theorems, or methods (Duan et al., 2025; Hao et al., 2022). For efficient and accurate annotation, we distill labeled data from DeepSeek-R1 (DeepSeek-AI et al., 2025) and fine-tune Qwen2.5-14B-Instruct (Qwen et al., 2025) to serve as our extractor (see Appendix D.1). We further consolidate the extracted KPs (see Appendix B.1), resulting in a final set of 10M high-quality KPs. Notably, 75.43% of QA instances examine multiple KPs, indicating strong interrelations among KPs and motivating us to construct a KP graph with strong edge associations.

Human Evaluation of KP Quality. To validate that extracted KPs capture genuine conceptual knowledge rather than superficial keywords, we conducted a human evaluation with domain experts. Annotators assessed whether KPs correctly represent the core concepts examined by each QA instance, achieving an acceptability rate of **97.5%**. This confirms that our extraction pipeline produces

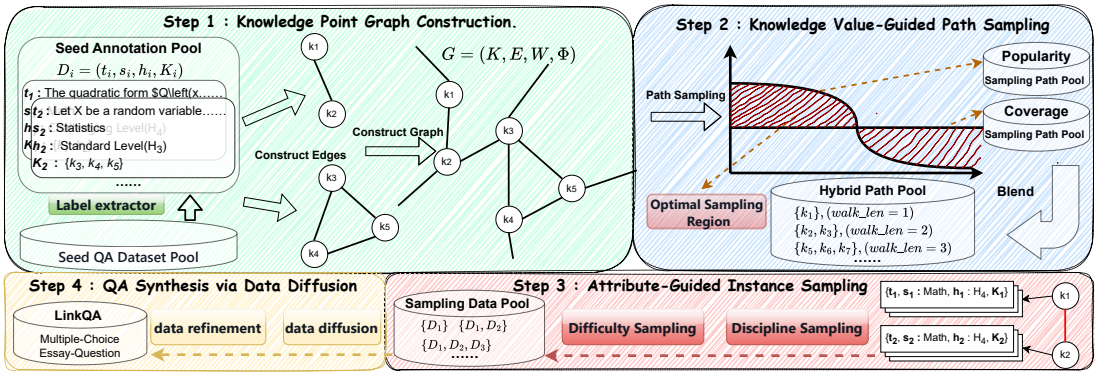

Figure 2: An overview of the LinkSyn pipeline. **Step 1: Knowledge Point Graph Construction** — We construct the KP graph based on the co-occurrence of KPs in the seed data. **Step 2: Knowledge Value-Guided Path Sampling** — Two sampling policies are utilized to balance KP coverage and popularity. **Step 3: Attribute-Guided Instance Sampling** — Instances are sampled by controlling the distributions of difficulty and discipline.

meaningful conceptual units. Furthermore, the fine-tuned Qwen-based extractor achieves an 80.8% agreement rate (edit distance ≤ 3) with DeepSeek-R1 annotations on held-out data, demonstrating reliable distillation quality. The two-stage KP consolidation (Appendix B.1) further merges near-duplicate descriptions into 10M atomic KPs, ensuring each KP represents a distinct, well-defined concept.

Knowledge Point Graph Construction. The KP graph is built from the annotated QA dataset $A = \{D_i\}_{i=1}^{m}$, where each item $D_i = (t_i, s_i, h_i, K_i)$ consists of the question text instance t_i, the discipline s_i, the difficulty h_i, and its associated KP set $K_i = \{k_{i1}, k_{i2}, \ldots, k_{in_i}\}$. We construct the KP graph $G = (K, E, W, \Phi)$ as follows: $K = \bigcup_{i=1}^{m} K_i$ is the set of unique KPs; E is the set of undirected edges, where an edge e_{k_p,k_q} exists if k_p and k_q co-occur in any instance:

$$E = \{\, e_{k_p,k_q} \mid k_p, k_q \in K,\ k_p \neq k_q,\ \exists D_l \in A \text{ s.t. } \{k_p, k_q\} \subseteq K_l \,\}$$

$W : E \to \mathbb{N}^+$ is the edge weight function, denoting the number of instances in which k_p and k_q co-occur:

$$W(e_{k_p,k_q}) = |\{D_l \in A \mid k_p, k_q \in K_l\}|$$

$\Phi : K \to 2^A$ maps each KP to the set of original data where it appears:

$$\Phi(k) = \{D_l \in A \mid k \in K_l\}$$

Graph Analysis. The resulting KP graph encompasses 10M nodes interconnected by 153M edges. Connectivity analysis reveals one giant connected component containing over 92% of texts and more than 89% of KPs (diameter 29), alongside numerous smaller components each confined to single-discipline domains. Community detection via the Leiden algorithm yields 21,806 KP clusters, with the dominant subject in each cluster averaging 86.76% of content, demonstrating that KPs naturally aggregate along disciplinary boundaries while maintaining cross-domain connections. Both node degree and edge weight distributions exhibit pronounced long-tail characteristics, motivating our hybrid sampling strategy below (detailed analysis in Appendix B.2).

2.3 Knowledge Value-Guided Path Sampling

Task Formulation. Based on the KP graph G, we design various random walk sampling policies $\mathcal{P}$ to obtain a set $\Pi = \{\pi_i\}_{i=1}^{M}$ containing M paths. Each path $\pi_i = \{k_{i1}, k_{i2}, \ldots, k_{il}\}$ consists of l sequentially connected KPs. We deliberately set $l \in \{1, 2, 3\}$ to control the complexity of the generated questions.

Design Objectives. We aim for the resulting path set to simultaneously satisfy two objectives: (1) the sampled connections between knowledge points should remain consistent with the knowledge association strength reflected in the seed corpus (Qin et al., 2025), thereby prioritizing the preservation of reliable logical relationships; (2) the sampling process needs to possess sufficient explorability (Huang et al., 2025), allowing weak connections and peripheral regions to be effectively reached, which enhances the coverage of long-tail knowledge and the diversity of combinations. Since the

edge weights of the knowledge point graph are constructed from the co-occurrence statistics of knowledge points in the seed data, the proportion of edge weights within a local neighborhood naturally characterizes *the seed-induced knowledge association structure*. We formalize these two objectives within the local neighborhood of the graph and derive a controllable sampling trade-off.

Let the neighborhood of a knowledge point k be $N(k) = \{k' \mid e_{k,k'} \in E\}$, and its weighted degree be $d_w(k) = \sum_{x \in N(k)} W(e_{k,x})$. We define the local co-occurrence transition distribution induced by edge weights: for any $k' \in N(k)$, let

$$Q(k' \mid k) = \frac{W(e_{k,k'})}{d_w(k)}, \quad k' \in N(k). \quad (1)$$

This distribution characterizes the relative co-occurrence strength proportions of various neighbors with respect to k within the local neighborhood centered at k, which can be viewed as the empirical association structure of the seed corpus on the graph.

First, to achieve the *structural alignment* objective, we require the strategy to reproduce the aforementioned co-occurrence proportion structure within each local neighborhood as much as possible, i.e., making $P(\cdot \mid k)$ close to $Q(\cdot \mid k)$. We use KL divergence to measure this degree of deviation:

$$\mathcal{L}_1(P) = \sum_{k \in K} \rho(k) \mathrm{KL}(P(\cdot \mid k) \parallel Q(\cdot \mid k)), \quad (2)$$

where $\rho(k)$ is the node weight, used to aggregate alignment errors across different neighborhoods.

Second, to achieve the *long-tail exploration* objective, we hope to maintain higher explorability within each local neighborhood, avoiding the sampling process from concentrating solely on a few high-density regions along strong co-occurrence edges. To this end, we define the uniform distribution over the neighborhood as

$$U(k' \mid k) = \frac{1}{|N(k)|}, \quad k' \in N(k), \quad (3)$$

and characterize the strategy's deviation from *uniform exploration* using the same KL form:

$$\mathcal{L}_2(P) = \sum_{k \in K} \rho(k) \mathrm{KL}(P(\cdot \mid k) \parallel U(\cdot \mid k)). \quad (4)$$

By unifying structural alignment and long-tail exploration into a single framework, we obtain the following controllable trade-off objective:

$$\min_P (1-\lambda)\mathcal{L}_1(P) + \lambda \mathcal{L}_2(P), \quad \lambda \in [0,1]. \quad (5)$$

Hybrid Sampling Policy. To optimize the objective in Eq. 5, we adopt a direct path-level mixing strategy. We independently sample paths using the coverage-priority policy p^a (corresponding to U) and the popularity-priority policy p^b (corresponding to Q), and blend them directly using the trade-off parameter λ:

$$\Pi_{\text{hybrid}} = \lambda \cdot \Pi_{p^a} + (1-\lambda) \cdot \Pi_{p^b} \quad (6)$$

where $\lambda \in [0,1]$ controls the balance between exploration and structural alignment. We provide a theoretical analysis in Appendix B.3, proving that this linear mixing of paths induces the optimal distribution required by the Knowledge Value objective.

2.4 Attribute-Guided Instance Sampling

Task Formulation. Given the set of sampled KP paths Π, we construct seed datasets $\mathcal{S} = \{S_\pi \mid \pi \in \Pi\}$, where

$$S_\pi = \{t_i \mid D_i = (t_i, s_i, h_i, K_i) \in \Phi(k_i), k_i \in \pi\}$$

Each text t_i is sampled from $\Phi(k_i)$ and is required to satisfy the target attribute distributions.

Distribution Constraints. We specify two attribute distributions: discipline and difficulty. Regarding difficulty, the original corpus contains a limited proportion of high-difficulty data (see Appendix A.3). To rectify this imbalance and improve performance on challenging problems (Tong et al., 2024), we sample difficulty levels with the following probabilities: 10% for H1, 15% for H2, and 25% for each of H3, H4, and H5, representing a gradient from easiest to hardest. For discipline, we focus on mathematics to create the $\text{LinkQA}_{\text{Math}}$ subset, which facilitates targeted evaluation of mathematical reasoning (Huang et al., 2024).

Difficulty and Discipline Annotation. We categorize disciplines into 62 categories and calibrate difficulty levels across five scales. We fine-tune Qwen2.5-7B-Instruct distilled from DeepSeek-R1 for large-scale automated labeling (detailed in Appendix D.1).

Instance Selection. For each node k_i in the sampled path, we select a supporting instance t^* from the set $\Phi(k_i)$ with difficulty h_D closest to the target difficulty h, prioritizing those with matching discipline s:

$$t^* = \underset{t \text{ in } D \in \Phi(k_i), s_D = s \text{ if possible}}{\arg\min} |h_D - h| \quad (7)$$

Algorithm 1 KP Path Sampling and Seed Instance Selection

Input: KP graph $G = (K, E, W, \Phi)$; sampling policies p^a and p^b; mixing parameter α; difficulty distribution ρ_h; discipline distribution ρ_s; path length l; path sample number M

Output: Sampled instances $\mathcal{S}$

Function PS(G, p, l, M): // Path Sampling
  Initialize $\Pi \leftarrow \emptyset$ // Set of sampled paths
  while $|\Pi| < M$ **do**
    $k_1 \sim p(k_1)$ on K; $\pi \leftarrow [k_1]$
    for $t = 1$ to $l-1$ **do if** $N(k_t) \neq \emptyset$ **then** sample k_{t+1} from $N(k_t)$ with $p(k_t, k_{t+1})$; append k_{t+1} to π
    if $\pi \notin \Pi$ **then** add π to Π
  return Π
$\Pi^a \leftarrow$ PS(G, p^a, l, M); $\Pi^b \leftarrow$ PS(G, p^b, l, M)
$\Pi_{\text{hybrid}} = \alpha \cdot \Pi_{p^a} + (1-\alpha) \cdot \Pi_{p^b}, \quad \alpha \in [0, 1]$
Initialize $\mathcal{S} \leftarrow \emptyset$ // Set of sampled instances
for each path π in Π_{hybrid}
  Initialize $\mathcal{S}_\pi \leftarrow \emptyset$; Sample $h \sim \rho_h$, $s \sim \rho_s$
  for each node k_t in π
    Sample t^* according to Eq. (7) with k_t, h, and s from instances not in $\mathcal{S}_\pi$
    Add t^* to $\mathcal{S}_\pi$
  if $\mathcal{S}_\pi \notin \mathcal{S}$ **then** add $\mathcal{S}_\pi$ to $\mathcal{S}$
return $\mathcal{S}$

2.5 QA Synthesis via Data Diffusion

The algorithm for obtaining related seed sets via path and instance sampling is described in Algorithm 1. We sample 20M seed groups for each combination of random walk length $l \in \{1, 2, 3\}$ and sampling policy, and additionally perform mathematics-constrained sampling for the LinkQA$_{\text{Math}}$ subset, and then blend them in equal proportions ($\alpha = 0.5$). Utilizing these seed data, we employ DeepSeek-R1 for QA synthesis based on a controlled comparison with Qwen2.5-72B-Instruct (Appendix C), and DeepSeek-V3 (DeepSeek-AI et al., 2024) for answer refinement (detailed in Appendix D.2). Subsequently, we perform comprehensive data cleaning and decontamination. To ensure evaluation integrity, we apply a two-stage decontamination pipeline against all 9 evaluation benchmarks: (1) 10-gram matching removes any synthesized QA pair sharing a contiguous 10-gram with any benchmark test instance; (2) embedding cosine similarity filtering using Sentence-T5 (Ni et al., 2021) embeddings removes pairs exceeding a semantic similarity threshold, providing a safety net beyond surface n-gram matching (Shao et al., 2024). Approximately **1%** of synthesized data is removed; inspection reveals that filtered samples predominantly share common entities (e.g., the same person or scientific concept) with benchmark instances rather than genuine duplication of test problems, indicating minimal actual contamination risk. Low-quality data is also filtered (see Appendix A.5). This rigorous pipeline ultimately yields the high-quality 50B-token LinkQA dataset. Finally, we blend LinkQA with high-quality corpora, KnowEdu, to construct the training dataset. KnowEdu is curated from pre-training corpora, where the QuRater (Wettig et al., 2024) quantifies knowledge density and the educational classifier from FineWeb-Edu (Penedo et al., 2024) evaluates educational utility. Texts rated highly in both knowledge density and educational utility are retained to form KnowEdu.

3 Experiments

3.1 Experimental Setup

Training Details. The effectiveness of LinkQA is validated during continual pre-training using Llama-3 8B, which is pre-trained on 10T tokens. Our main experiments commence continual pre-training from the 2T-token checkpoint using a 1:1 mixture of QA and KnowEdu, with 40B tokens, implemented via the Megatron framework (Narayanan et al., 2021) and optimized by the Adam algorithm. The training employs a linearly decaying learning rate schedule initialized at 1.9×10^{-4} and terminating at 1.9×10^{-5}. Further scaling experiments systematically examine the model size scale by evaluating 1.7B and 16B architectures under identical 40B-token configurations, and initial FLOPs scale of 2T and 10T tokens for the 8B model. Details of the training setup are provided in Appendix A.1.

Evaluation. We adopt 9 benchmarks for comprehensive evaluation. Knowledge-intensive benchmarks include MMLU (Hendrycks et al., 2021b), CMMLU (Li et al., 2024), C-Eval (Huang et al., 2023), MMLU-Pro (Wang et al., 2024), and MMLU-STEM. Reasoning abilities are measured using WinoGrande (Sakaguchi et al., 2021), BBH (Suzgun et al., 2023), ARC-C (Clark et al., 2018), and DROP (Dua et al., 2019).

Dataset	MMLU	CMMLU	C-Eval	M-Pro	STEM	W.G.	BBH	ARC-C	DROP	AVG.
Pre-training	55.08	52.23	57.11	24.32	45.17	51.50	35.79	70.50	42.31	48.22
FineWeb-Edu	56.23	58.88	56.80	25.46	47.78	53.50	34.38	69.60	39.44	49.12
KnowEdu	58.17	62.99	61.98	25.64	49.16	54.50	35.12	71.50	41.07	51.13
Nemotron-CC	58.62	62.02	59.02	27.68	49.25	55.00	34.86	73.00	41.63	51.23
YulanQA	56.15	58.95	57.53	24.89	47.09	53.00	35.75	73.00	42.18	49.84
Nemotron-MIND	58.48	61.11	60.98	30.32	51.55	52.00	37.69	73.00	46.33	52.38
MegaMathQA	55.98	59.04	58.91	26.86	48.59	55.50	35.64	68.50	43.31	50.26
JiuZhang3.0	56.55	60.30	59.52	27.43	48.68	55.00	36.33	71.50	45.05	51.15
LinkQA	**63.98**	**66.35**	**65.59**	**30.57**	**56.95**	**56.50**	**38.09**	**79.50**	**49.94**	**56.39**

Table 1: Comparison across benchmarks. The best is in bold. Abbreviations: M-Pro = MMLU-Pro, STEM = MMLU-STEM, W.G. = WinoGrande.

Dataset	Synth.	CoT	Type	Domain	Tokens
N-CC QA	✔	✗	QA	General	51B
YulanQA	✗	✔	QA	General	4.92B
N-MIND	✔	✔	Conv.	Math	138B
MegaMathQA	✔	✔	QA	Math	7.0B
JiuZhang3.0	✔	✔	QA	Math	4.6B
LinkQA	✔	✗	QA	General	30B
$\text{LinkQA}_{\text{CoT}}$	✔	✔	QA	General	50B

Table 2: Comparison with large-scale QA datasets. $\text{LinkQA}_{\text{CoT}}$ extends LinkQA with supplementary math CoT data $\text{LinkQA}_{\text{MathCoT}}$.

Baselines. We employ two baseline evaluation paradigms. **The first assesses 40B-token general corpora**, comprising the standard pre-training dataset and the web-sourced educational corpora FineWeb-Edu (Penedo et al., 2024), alongside KnowEdu, our curated high-quality knowledge-rich and educational data. **The second paradigm assesses 40B-token QA blend following a 1:1 mixing ratio between KnowEdu and QA datasets.** General baselines include QA subset of Nemotron-CC (Su et al., 2025), and YulanQA, a QA subset extracted from the continual pre-training dataset of Chen et al. (2025). Mathematical baselines incorporate Nemotron-MIND (Akter et al., 2025), a dataset of synthetic math dialogues; MegaMathQA, a QA subset derived from MegaMath-Synthetic (Zhou et al., 2025); and JiuZhang3.0 (Zhou et al., 2024), a dataset of structured math problems with chain-of-thought (CoT). A comparison of all datasets is provided in Table 2. For Nemotron-CC, we maintain its original 9:1 document-to-QA ratio and substitute Nemotron-CC Document with KnowEdu, reporting the optimal configuration.

3.2 Main Results

The main experimental results are shown in Table 1, with mathematical results in Table 3. We find that:

LinkQA achieves significant superiority over the general corpus baselines, establishing SOTA performance. Compared to the pre-training baseline, LinkQA achieves an average improvement of **11.51%** on MMLU and CMMLU, and 8.17% across all benchmarks. LinkQA also demonstrates advantages over FineWeb-Edu and KnowEdu, with average improvements of 7.27% and 5.26%, respectively. For the average performance across all benchmarks, LinkQA outperforms the suboptimal Nemotron-MIND by 4.01%, demonstrating the advantage of our LinkSyn method. Since all QA baselines are mixed with KnowEdu under the identical 1:1 protocol, the gap is attributable to QA data quality rather than the mixing effect (KnowEdu alone reaches 51.13%; LinkQA+KnowEdu reaches 56.39%).

LinkQA demonstrates sustained leading advantages during training. Figure 1 reveals an expanding performance gap between LinkQA and the baselines as training progresses, particularly evident at 20B tokens. This indicates that LinkQA can continuously provide high-quality knowledge signals to models, promoting knowledge accumulation and integration while delivering long-term capability enhancement.

For mathematical tasks, $\text{LinkQA}_{\text{MathCoT}}$ improves by 7.07% on average over the strongest baseline and achieves SOTA average performance. As presented in Table 3, $\text{LinkQA}_{\text{MathCoT}}$ outperforms JiuZhang3.0 by 4.85% on GSM8K. $\text{LinkQA}_{\text{Math}}$ demonstrates best performance on mathematical subsets of MMLU and significantly surpasses the second-best JiuZhang3.0 by 4.54%

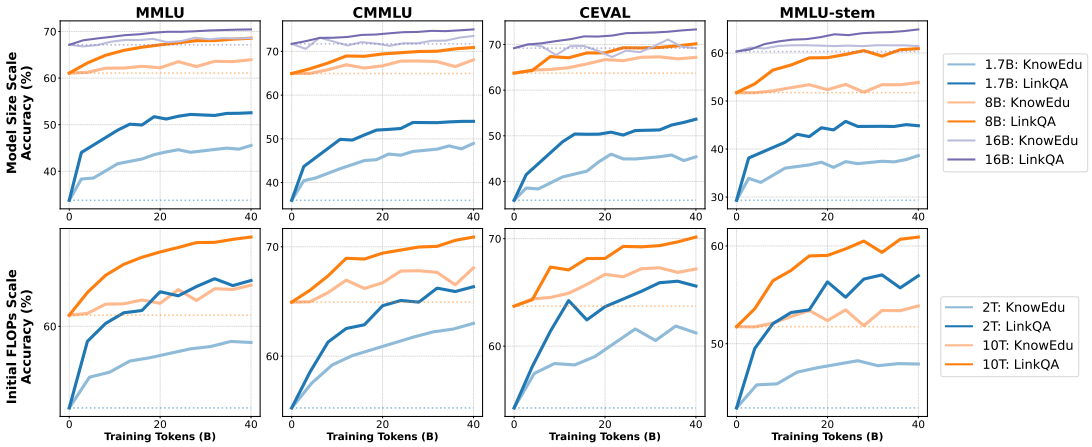

Figure 3: Scaling analysis of LinkQA across model and computational dimensions (detailed in Appendix A.2). For model size scalability, we use initial checkpoints of 4T, 10T, and 10T tokens for 1.7B, 8B, and 16B models.

Dataset	CoT	G.	M.	Elem.	High.	Coll.	AVG.
Pre-training	-	33.95	6.50	36.50	30.50	25.00	26.49
FineWeb-Edu	-	31.49	2.50	38.00	31.00	30.00	26.60
KnowEdu	-	32.56	8.00	37.50	27.50	31.00	27.31
N-MIND	✔	47.42	13.50	43.50	29.50	37.00	34.18
MegaMathQA	✔	44.65	6.50	47.00	31.50	35.00	32.93
JiuZhang3.0	✔	56.27	**23.00**	42.50	30.50	31.00	36.65
LinkQA	✗	39.41	9.50	44.50	38.00	44.00	35.08
$\text{LinkQA}_{\text{Math}}$	✗	42.96	13.50	**57.00**	**46.50**	**46.00**	41.19
$\text{LinkQA}_{\text{MathCoT}}$	✔	**61.12**	21.50	53.50	44.50	38.00	**43.72**

Table 3: Comparison of mathematical performance. The best and second best are in bold and underlined, respectively. Abbreviations: G. = GSM8K, M. = MATH, Elem. = MMLU: elementary-mathematics, High. = MMLU: high-school-mathematics, Coll. = MMLU: college-mathematics.

Dataset	MMLU	C-Eval	ARC-C	DROP	AVG.
$\lambda = 1$	59.40	61.43	71.50	44.48	59.20
$\lambda = 0$	58.96	61.40	73.50	46.04	59.98
$\lambda = 0.5$	59.94	62.89	74.00	47.01	60.96
$l = 1$	58.91	61.79	72.00	44.30	59.25
$l = 2$	59.73	62.46	73.50	46.89	60.65
$l = 3$	59.88	63.04	74.50	46.92	61.09

Table 4: Comparison of different sampling policies ($\lambda \in \{0, 0.5, 1\}$) and different walk lengths ($l \in \{1, 2, 3\}$).

on average. These highlight the effectiveness of LinkQA for comprehensive mathematical tasks. Note that the 50B-token LinkQA consists of 30B pure QA tokens and 20B CoT math tokens ($\text{LinkQA}_{\text{MathCoT}}$). The main CPT experiments use 20B QA tokens + 20B KnowEdu = 40B tokens; the CoT portion is used separately in the math experiments above, achieving SOTA results. In total, 40B out of 50B tokens have been used and validated.

3.3 Scaling Analysis

We conduct scaling analysis experiments to validate the robustness of LinkQA, shown in Figure 3.

LinkQA brings significant performance improvements across different model sizes. For models with sizes of 1.7B, 8B, and 16B, as training progresses, the accuracy of LinkQA across different benchmarks consistently outperforms that of KnowEdu. This consistent enhancement across various model sizes demonstrates the excellent quality and generalization capability of LinkQA.

LinkQA consistently improves model performance regardless of initial FLOPs scale. Different initial FLOPs reflect varying initial model capability at the start of continual pre-training. At both the 2T-token checkpoint and the 10T-token checkpoint, LinkQA demonstrates superior performance compared to KnowEdu. As training progresses, LinkQA exhibits an upward trend. This consistent performance across different pre-training stages with varying initial model capabilities further validates the high-quality characteristics of LinkQA.

3.4 Ablation Studies

For sampling policy ablation, we test $\lambda = \{1, 0.5, 0\}$ (Eq. 6), while maintaining fixed ratios of each $l \in \{1, 2, 3\}$. For random walk length ablation, we fix $\lambda = 0.5$ and synthesize datasets using exclusively $l \in \{1, 2, 3\}$ (Section 2.3). In both experiments, we combine the 4B-token LinkQA with 12B-token KnowEdu.

For sampling policies, the hybrid $\lambda = 0.5$ achieves superior performance. As shown in Table 4, $\lambda = 0.5$ improves by 1.76% compared to $\lambda = 1$ and by 0.98% compared to $\lambda = 0$ on average, verifying the effectiveness of combining coverage and popularity sampling. For pure strategy comparison, in knowledge-intensive tasks such as MMLU and C-Eval, $\lambda = 1$ shows a relative advantage over $\lambda = 0$, indicating that KP coverage is more important for this type of task. Conversely, in complex reasoning tasks such as ARC-C and DROP, $\lambda = 0$ is preferable, emphasizing the importance of knowledge popularity for such tasks.

For random walk length, increasing the length consistently improves the performance. As shown in Table 4, the $l = 3$ random walk achieves an average improvement of 1.84% over $l = 1$ and 0.44% over $l = 2$. This demonstrates that increasing random walk length in KP graph sampling effectively enhances the diversity and coverage of synthesized training data, leading to notable performance improvements, particularly for benchmarks focused on complex and compositional reasoning. However, it is important to note that longer walks also increase data synthesis costs, emphasizing the need for a practical trade-off between diversity and efficiency in large-scale data generation.

Mixture ratio sensitivity. Beyond the 1:1 QA/-KnowEdu ratio used throughout, we additionally tested 1:2 and 1:3; the relative ranking of QA datasets is consistent across all ratios, in line with the stability of CPT-based evaluation (Guo et al., 2024).

3.5 Synthetic Data Analysis

We present a multi-dimensional analysis of synthesized data, including semantic diversity and distribution analysis. For each setting ($l = 1, 2, 3$), we sample 10,000 groups (S_q, G_q), where $S_q = \{s_{q_1}, \ldots, s_{q_k}\}$ ($k = 1, 2, 3$) are seed data, and $G_q = \{g_{q1}, \ldots, g_{qm}\}$ ($m = 10$) are the corresponding generated data. To establish meaningful comparisons, we create control groups using randomly paired seeds (random-2/3-seed) without KP graph guidance. All results are visualized in Figures 4, 5, and 6.

Multi-seed synthesis achieves broader and more uniform semantic diffusion than single-seed generation. We use Sentence-T5 (Ni et al., 2021) to embed the sampled data. First, we compute the offset vectors of the generated data relative to their seed data. Figure 4 shows the t-SNE visualization of 500 such offsets per group, demonstrating that 2/3-seed distributions cover a larger and more uniform semantic space. Next, we calculate the mean cosine similarity among the generated data within each group: $\text{mean}_q \, \text{mean}_{i \neq j} \, \cos(g_{qi}, g_{qj})$. As shown in the left panel of Figure 5, using more seeds results in lower similarity and thus greater diversity among the generated data.

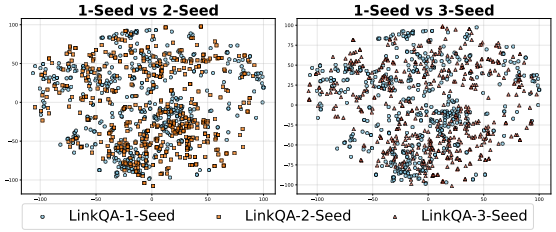

Figure 4: t-SNE visualization of semantic offsets between generated QA and seed data embeddings

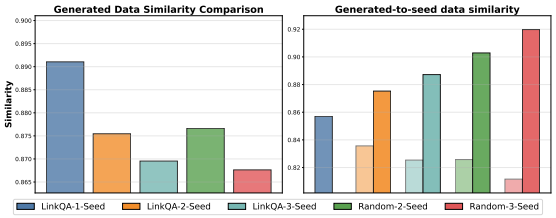

Figure 5: Left: Mean pairwise similarity among generated QA. Right: Minimum and maximum similarity between generated and seed data.

KP graph-based sampling enables effective semantic fusion across seeds. To evaluate semantic integration, we compute $\text{mean}_q \, \text{mean}_i \, \text{agg}_{s \in S_q} \, \text{sim}(g_{qi}, s)$, where agg is either max or min, representing the maximum or minimum similarity between each generated data and its corresponding seeds. As shown in the right panel of Figure 5, graph-based multi-seed generation yields a much smaller gap (0.04/0.06) than random sampling (0.07/0.11), with comparable overall diversity. This indicates that graph-based sampling effectively fuses semantics across seeds, while random sampling produces examples closely tied to a single seed.

LinkQA contains a higher proportion of high-difficulty data than baselines, and multi-seed synthesis further increases this proportion. The left panel of Figure 6 illustrates the difficulty distributions of LinkQA and baselines, based on 100,000 randomly sampled and annotated in-

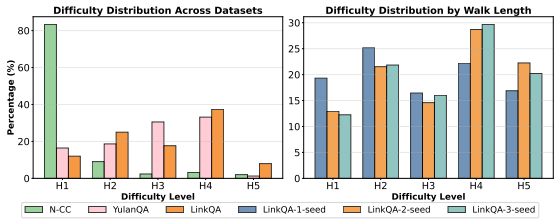

Figure 6: Left: Difficulty distribution of LinkQA vs. baseline. Right: Difficulty of 1/2/3-seed generated data with uniform seed difficulty.

stances from each dataset. LinkQA includes approximately $10\times$ more high-difficulty items (H4, H5) than Nemotron-CC. Notably, it even surpasses YulanQA, which, although not synthetic, is filtered for challenging QA yet still contains fewer high-difficulty items. The right panel demonstrates that multi-seed synthesis further elevates the proportion of challenging questions, with 2/3-seed methods yielding 10% more high-difficulty items than the 1-seed method, indicating that integrating multiple knowledge sources enhances question complexity. Detailed distribution analysis of LinkQA is provided in Appendix A.4&A.5. We further conduct case study of LinkQA in Appendix E.

Quality review. To assess the quality of LinkQA, we randomly sample 100 QA pairs from each difficulty level and have a professional annotator evaluate solvability and answer accuracy. The correctness rate is 98% for H1/H2, 94% for H3, and 87% for H4/H5, confirming that the majority of synthetic QA pairs meet the correctness criterion. As a further safeguard, the DeepSeek-V3 refinement pipeline (Appendix D.2) re-derives answers, retaining $\sim$84% as verified, correcting $\sim$10%, and discarding $\sim$6% as unsolvable or malformed. We further conduct case study of LinkQA in Appendix E.

4 Related Work

Existing work on pre-training data synthesis for LLMs has produced a diverse range of corpora and methods. General corpora such as FineWeb-Edu (Penedo et al., 2024) provide broad coverage but lack explicit QA supervision. In contrast, QA data and curated chain-of-thought corpora show superior performance in mid-training (Wang et al., 2025; Zhang et al., 2026; Tu et al., 2025). Several studies aim to improve QA quality: Maini et al. (2024) rephrases pre-training corpora into QA form; Cheng et al. (2024) designs instruction-driven synthesis; and Jiang et al. (2025a) evaluates QA integration in continual pre-training. Large-scale pipelines include Nemotron (Su et al., 2025), which generates 499.5B-token of document–QA pairs; MIND (Akter et al., 2025), which creates 138B-token of role-specific math dialogues; MegaMath (Zhou et al., 2025), a 7B-token dataset refined from mathematics-related webpages; and JiuZhang3.0 (Zhou et al., 2024), a 4.6B-token distilled corpus for mathematical QA. These methods show promise for synthetic QA but face challenges in scaling reasoning quality, ensuring concept diversity, and supporting multi-domain generalization required by modern large-scale MoE systems (Team et al., 2025a,b).

Recent work has increasingly emphasized controllable synthesis, motivating the use of explicit structural priors to steer data generation. Accordingly, graph-based sampling has been adopted to inject knowledge structure into synthesis. Entity-graph approaches (Qin et al., 2025; Jiang et al., 2025b) connect texts via co-occurring entities, but entities often capture surface mentions rather than the underlying concepts being examined. In contrast, we construct graphs over knowledge points to model tighter logical relations and to enable direct control over difficulty, discipline, and KP distributions, forming the basis of LinkSyn and its resulting dataset LinkQA.

5 Conclusion

In this paper, we introduce LinkSyn, a novel KP graph-based synthesis framework. By extracting KPs from QA seed data and constructing KP graphs, LinkSyn performs diffusion-based QA synthesis via DeepSeek-R1, based on multiple seeds that are strongly linked by KPs and sampled from graph walks. The synthesized LinkQA dataset significantly advances multi-disciplinary capabilities, as demonstrated by an 11.51% average improvement on MMLU and CMMLU when continually pre-training Llama-3 8B. These SOTA results, coupled with consistent gains across model size and initial FLOPs scales, underscore LinkSyn's efficacy in generating diverse, valuable synthetic QA data.

Limitations

This paper demonstrates that path sampling based on knowledge point graphs, combined with multi-seed diffusion synthesis, can enhance data diversity and knowledge coverage while maintaining at-

tribute controllability, such as subject and difficulty. All experiments use Llama-3 variants because mid-training (continual pre-training) requires a checkpoint taken *before* the final annealing phase of pre-training. Open-source model families (Qwen, Gemma, etc.) typically only release their final annealed checkpoint, making them unsuitable for CPT experiments. Despite this constraint, we validate across three model sizes (1.7B, 8B, 16B) and two initial FLOPs scales (2T, 10T), covering a $10\times$ range in model size and $5\times$ in training compute. Furthermore, as the model continuously evolves into new states during the training process, how to extend the path sampling and synthesis pipeline into a dynamic, multi-stage closed-loop mechanism—for instance, by periodically re-evaluating sampling trade-off coefficients and path lengths to adjust synthesis ratios—remains a subject for further systematic research. While the current framework primarily provides control at the levels of subject, difficulty, and knowledge point association structures, future work could integrate more fine-grained dimensions, such as question types, reasoning formats, cross-knowledge dependency depth, and compositional intensity, into a unified constraint system to achieve more precise distribution regulation and stable generalization gains. Finally, although we have validated the scalability of this method in large-scale settings, it remains a worthwhile direction to explore how to summarize stable patterns across a broader range of model families and domain scenarios, while further reducing the costs of synthesis and verification under quality constraints.

References

Syeda Nahida Akter, Shrimai Prabhumoye, John Kamalu, Sanjeev Satheesh, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. MIND: Math informed synthetic dialogues for pretraining LLMs. In *The Thirteenth International Conference on Learning Representations*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367.

Yaoyao Chang, Lei Cui, Li Dong, Shaohan Huang, Yangyu Huang, Yupan Huang, Scarlett Li, Tengchao Lv, Shuming Ma, Qinzheng Sun, Wenhui Wang, Furu Wei, Ying Xin, Mao Yang, Qiufeng Yin, and Xingxing Zhang. 2024. Redstone: Curating general, code, math, and qa data for large language models. *Preprint*, arXiv:2412.03398.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Xin Zhao, Zhicheng Dou, Jiaxin Mao, Yankai Lin, Ruihua Song, Jun Xu, Xu Chen, Rui Yan, Zhewei Wei, Di Hu, Wenbing Huang, and Ji-Rong Wen. 2025. Towards effective and efficient continual pre-training of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Feiyu Duan, Xuemiao Zhang, Sirui Wang, Haoran Que, Yuqi Liu, Wenge Rong, and Xunliang Cai. 2025. Enhancing llms via high-knowledge data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23832–23840.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang,

Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. 2024. Efficient continual pre-training by mitigating the stability gap. *Preprint*, arXiv:2406.14833.

Fei Hao, Yanqi Gong, Wangyang Yu, and Vincenzo Loia. 2022. Knowledge points navigation based on three-way concept lattice for autonomous learning. *Pattern Recognition Letters*, 163:96–103.

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

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

Binyuan Huang, Yuqing Wen, Yucheng Zhao, Yaosi Hu, Yingfei Liu, Fan Jia, Weixin Mao, Tiancai Wang, Chi Zhang, Chang Wen Chen, Zhenzhong Chen, and Xiangyu Zhang. 2024. Subjectdrive: Scaling generative data in autonomous driving via subject control. *Preprint*, arXiv:2403.19438.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, page 1–55.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jinhao Jiang, Junyi Li, Xin Zhao, Yang Song, Tao Zhang, and Ji-Rong Wen. 2025a. Mix-CPT: A domain adaptation framework via decoupling knowledge learning and format alignment. In *The Thirteenth International Conference on Learning Representations*.

Xuhui Jiang, Shengjie Ma, Chengjin Xu, Cehao Yang, Liyu Zhang, and Jian Guo. 2025b. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models. *Preprint*, arXiv:2505.00979.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *Preprint*, arXiv:1607.06275.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2025. Scaling data-constrained language models. *Preprint*, arXiv:2305.16264.

Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, page 1–1.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA. Association for Computing Machinery.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. 2025. Scaling laws of synthetic data for language models. *Preprint*, arXiv:2503.19551.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Meituan LongCat Team, Bayan, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, Chengcheng Han, Chenguang Xi, Chi Zhang, Chong Peng, Chuan Qin, Chuyu Zhang, Cong Chen, Congkui Wang, and 163 others. 2025a. Longcat-flash technical report. *Preprint*, arXiv:2509.01322.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chao Zhang, Chengcheng Han, Chenhui Yang, Chi Zhang, Chong Peng, Chuyu Zhang, Cong Chen, Fengcun Li, Gang Xu, Guoyuan Lin, Hao Jiang, Hao Liang, and 108 others. 2025b. Introducing longcat-flash-thinking: A technical report. *Preprint*, arXiv:2509.18883.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. DART-math: Difficulty-aware rejection tuning for mathematical problem-solving. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. 2025. A survey on llm mid-training. *Preprint*, arXiv:2510.23081.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? limits of llm scaling based on human-generated data. *Preprint*, arXiv:2211.04325.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025. Octothinker: Mid-training incentivizes reinforcement learning scaling. *Preprint*, arXiv:2506.20512.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *International Conference on Machine Learning*, pages 52915–52971.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2025. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*.

Xuemiao Zhang, Can Ren, Chengying Tu, Rongxiang Weng, Shuo Wang, Hongfei Yan, Jingang Wang, and Xunliang Cai. 2026. Expanding reasoning potential in foundation model by learning diverse chains of thought patterns. *Preprint*, arXiv:2509.21124.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P. Xing. 2025. Megamath: Pushing the limits of open math corpora. *Preprint*, arXiv:2504.02807.

Kun Zhou, Beichen Zhang, jiapeng wang, Zhipeng Chen, Xin Zhao, Jing Sha, Zhichao Sheng, Shijin Wang, and Ji-Rong Wen. 2024. Jiuzhang3.0: Efficiently improving mathematical reasoning by training small data synthesis models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Experimental Details

A.1 Training Details

The seeds are sourced through an extensive collection from open-source QA datasets, such as WebQA (Li et al., 2016), MathQA (Amini et al., 2019), and RedStone (Chang et al., 2024), extraction of QA data from web pages like Stack Overflow, as well as a portion acquired through procurement.

We employ DeepSeek-R1 for data synthesis and DeepSeek-V3 for answer refinement, with both models setting temperature to 0.6, top-p value to 0.95, and top-k to -1. All computations are executed on a dedicated cluster of 300 H20-141G GPUs.

We use 256 Ascend 910B NPUs to continually pre-train the Llama-3 8B model from the 2T-token checkpoint using 40B tokens of QA blend with KnowEdu, each model taking over 22 hours. We implement it via the Megatron framework (Narayanan et al., 2021), optimized by the Adam algorithm with standard $\beta_1 = 0.9$ and $\beta_2 = 0.95$ parameters. The training employs a global batch size of 960 and a linearly decaying learning rate schedule initialized at 1.9×10^{-4} and terminating at 1.9×10^{-5}.

In the model size scale experiment, we also test the performance of the dataset on the 1.7B and 16B models with the same settings as 8B. For the 1.7B model, we use 80 NPUs for training, each model taking over 38 hours. For the 16B model, we use 480 NPUs for training, each model taking over 21 hours. In Table A1, we present the model configuration of the 1.7B and 8B models.

We further analyze the computational cost and data scale for constructing 1M QA samples. Specifically, generating 1M QA pairs using DeepSeek-R1 requires 514.07 GPU hours on H20 GPUs, while answer refinement with DeepSeek-V3 costs an additional 318.72 GPU hours. For reference, 1M pure QA samples correspond to 0.093B tokens, and 1M CoT-augmented QA samples correspond to 0.437B tokens.

A.2 Scaling Details

The specific accuracy values of the final checkpoint in the scaling experiments are shown in Table A2.

A.3 Seed Data Distribution

The difficulty and discipline distribution of our seed data are illustrated in Figure A1. Regarding discipline distribution, the seed data covers all disciplines with balanced proportions, where mathematics accounts for the largest share at 25% but remains within reasonable bounds. However, the difficulty distribution shows significant imbalance, with 50% of the seed data concentrated at the H1 difficulty level. To address this imbalance, we implement difficulty control measures during the data sampling process.

Hyperparameter	1.7B	8B	16B
Precision	bfloat16	bfloat16	bfloat16
Layers	24	32	40
Hidden Size	2048	4096	5120
Attention Heads	32	32	64
Head Type	GQA	GQA	GQA
Intermediate Size	8192	14336	18432
Vocab Size	131072	131072	163840
Sequence Length	8192	8192	8192
Activation	SiLU	SiLU	SiLU
Position Embedding	RoPE	RoPE	RoPE

Table A1: Model structure of Llama-3.

Scale	Settings	MMLU	CMMLU	C-Eval	STEM
Model Size	1.7B: KnowEdu	45.32	48.95	45.39	38.63
	1.7B: LinkQA	52.56	54.00	53.63	44.85
	8B: KnowEdu	63.53	68.08	67.18	53.86
	8B: LinkQA	68.57	70.89	70.15	60.91
	16B: KnowEdu	68.61	72.53	69.23	61.44
	16B: LinkQA	70.45	75.03	73.30	64.91
Initial FLOPs	2T: KnowEdu	58.17	62.99	61.98	49.16
	2T: LinkQA	64.40	66.35	65.59	56.95
	10T: KnowEdu	63.53	68.08	67.18	53.86
	10T: LinkQA	68.57	70.89	70.15	60.91

Table A2: Accuracy of the final checkpoint in the scaling experiments. Abbreviations: STEM = MMLU-STEM. For model size scalability, we use initial checkpoints of 4T, 10T, and 10T tokens for 1.7B, 8B, and 16B parameter models, respectively.

A.4 LinkQA Distribution

The difficulty and subject distributions of our sampled seed data and LinkQA (we sampled 100k data for difficulty and subject annotation analysis) are shown in Figure A2. Our target difficulty distribution is (H1–H5, from easiest to hardest): 10%, 15%, 25%, 25%, 25%, but due to the scarcity of high-difficulty data, the actual difficulty can only approximate the target difficulty as described in Equation 7, resulting in some deviation in our final sampled seed data. Notably, LinkQA exhibits a higher proportion of high-difficulty questions compared to the seed data because multi-seed data synthesis tends to elevate the overall difficulty level. Regarding subject distribution, both datasets ap-

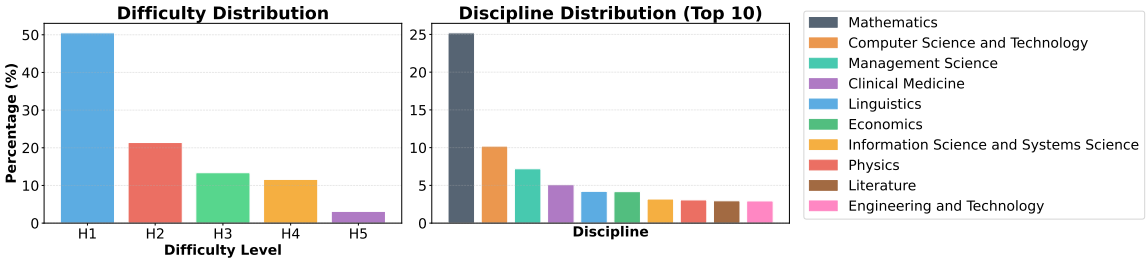

Figure A1: Distribution of all seed data: difficulty distribution (left) and discipline distribution (right).

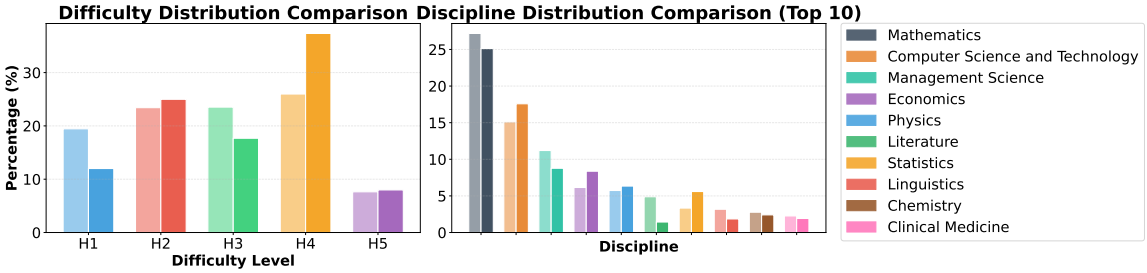

Figure A2: Distribution comparison of LinkSyn sampled seed data (light color) and LinkQA (dark color): difficulty distribution (left) and discipline distribution (right).

proximate natural distributions, though we observe that LinkQA shows reduced proportions in general subjects such as Mathematics and Literature compared to the sampled seed data. This occurs because knowledge points in these general subjects have more connections with other disciplines, leading to cross-disciplinary data appearing in the sampled knowledge point paths, which causes the subject distribution inconsistency between LinkQA and seed data.

A.5 Quality Review of LinkQA

To rigorously assess the quality of LinkQA, we randomly sample 100 QA pairs from each predefined difficulty level. A professional annotator evaluates each pair along two dimensions: (i) the solvability of the question, and (ii) the accuracy of the corresponding answer. A QA pair is deemed correct only if the question is solvable and the provided answer is fully accurate. The results, summarized in Table A3, demonstrate that the majority of synthetic QA pairs meet the correctness criterion. Moreover, the correctness rate exhibits a decreasing trend with increasing difficulty, indicating a correlation between difficulty level and quality metrics.

Difficulty Level	Correct (%)
H1/H2	98
H3	94
H4/H5	87

Table A3: Manual quality review results for LinkQA across different difficulty levels.

B Knowledge Point Graph

B.1 Knowledge Point Consolidation

As illustrated in Figure A3, we perform knowledge point (KP) consolidation in two stages. In the first stage, we standardize the case of all KPs, then group KPs by the first three identical characters (prefix length 3), and within each group, we cluster KPs such that the maximum pairwise edit distance in a cluster does not exceed the greater of 3 or $\text{int}(0.5 \times \max(\text{len}(s_1), \text{len}(s_2)))$. For each cluster, we use Qwen-14B to summarize and merge the KPs. In the second stage, we compute co-occurrence vectors for each KP and cluster those with cosine similarity above 0.9; again, we use Qwen-14B to summarize and consolidate each cluster. This detailed consolidation process improves the KP graph, but as it does not critically affect LinkSyn, therefore, the consolidation parameters are flexible and can be adjusted as needed. In our implementation, these steps result in a final set of 10M KPs.

B.2 Knowledge Point Graph Analysis

The knowledge point graph encompasses 10M nodes interconnected by 153M edges, exhibiting

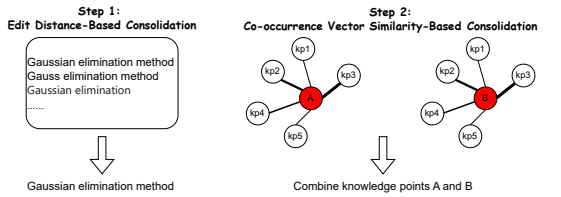

Figure A3: Two-step knowledge point consolidation process: edit distance-based deduplication (Step 1) and co-occurrence vector similarity-based deduplication (Step 2).

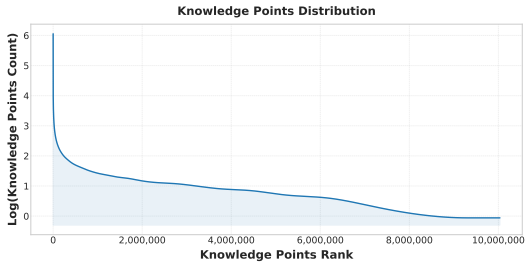

Figure A4: Knowledge points frequency distribution.

a notably sparse graph structure. As illustrated in Figure A4, both the text quantity distribution across nodes (ranging from 1 to 737,794) and edge weight distribution (ranging from 1 to 149,382) exhibit pronounced coverage characteristics. We performed connectivity analysis on the graph, revealing that our knowledge point network constructed from seed data consists of one giant connected component containing over 92% of texts and more than 89% of knowledge points (with a diameter of 29), alongside numerous smaller connected components. These smaller components contain fewer than 44 knowledge points each and are consistently confined to single discipline domains. The network's assortativity coefficient is merely 0.0892, indicating limited degree homophily. We apply the Leiden community detection algorithm on the graph and get 21,806 knowledge point clusters, with the dominant subject in each cluster averaging 86.76% of content, demonstrating that knowledge points naturally aggregate according to their disciplinary boundaries while maintaining crucial cross-domain connections.

B.3 Theoretical Justification of the Sampling Strategy

In this section, we provide a unified theoretical analysis showing that our hybrid path sampling strategy (Algorithm 1) optimizes the Knowledge Value trade-off objective in Eq. 5 by inducing (in expectation) the desired node visitation distribution.

Informal Summary. The key theoretical result is straightforward: when we mix paths from two policies—one that follows edge weights (popularity) and one that explores uniformly (coverage)—the resulting KP visitation distribution is exactly the unique minimizer of a trade-off objective that balances structural alignment with long-tail exploration. The mixing parameter λ directly controls this balance, and the approximation error is bounded by how well each base policy matches its target. The formal derivation below makes this precise.

Setup. Let K be the set of knowledge point (KP) nodes with $|K|$ elements, and let $\Delta^{|K|} = \{\nu \in \mathbb{R}_{\geq 0}^{|K|} : \sum_{k \in K} \nu(k) = 1\}$ denote the probability simplex. We use $\nu^{pop} \in \Delta^{|K|}$ for the seed-induced popularity distribution (denoted as Q in the main text), and $\nu^{cov} \in \Delta^{|K|}$ for the uniform coverage distribution (denoted as U). Throughout, $\lambda \in [0, 1]$ controls the trade-off.

Optimal Distribution Derived from the Objective. Recall the controllable trade-off objective in Section 2.3:

$$\min_{P} \mathcal{J}(P) = (1 - \lambda)\mathcal{L}_{align}(P) + \lambda \mathcal{L}_{explore}(P). \tag{8}$$

To obtain a closed-form target distribution, we consider a Euclidean relaxation of distribution matching:

$$\min_{\nu \in \Delta^{|K|}} \Big[(1 - \lambda)\|\nu - \nu^{pop}\|_2^2 + \lambda \|\nu - \nu^{cov}\|_2^2\Big]. \tag{9}$$

The objective in Eq. 9 is strictly convex in ν, hence admits a unique global minimizer. Introduce the Lagrangian:

$$\mathcal{L}(\nu, \eta) = (1 - \lambda) \sum_k (\nu(k) - \nu^{pop}(k))^2 + \lambda \sum_k (\nu(k) - \nu^{cov}(k))^2 + \eta\Big(\sum_k \nu(k) - 1\Big). \tag{10}$$

Taking derivative w.r.t. $\nu(k)$ and setting it to zero yields

$$2(1 - \lambda)\big(\nu(k) - \nu^{pop}(k)\big) + 2\lambda\big(\nu(k) - \nu^{cov}(k)\big) + \eta = 0. \tag{11}$$

Summing Eq. 11 over k and using $\sum_k \nu(k) = \sum_k \nu^{pop}(k) = \sum_k \nu^{cov}(k) = 1$ gives $\eta = 0$. Therefore, for each node k,

$$\nu^*(k) = (1 - \lambda)\nu^{pop}(k) + \lambda \nu^{cov}(k). \tag{12}$$

Since $\nu^{pop}, \nu^{cov} \in \Delta^{|K|}$ and $\lambda \in [0, 1]$, ν^* is a convex combination and thus also lies in $\Delta^{|K|}$.

Distribution Induced by Hybrid Path Mixing. We now characterize the node visitation distribution produced by Algorithm 1. Consider sampling M paths, each of length l. Let $N(k)$ denote the total number of visits to node k across all sampled paths, and define the induced (expected) visitation distribution as

$$\nu(k) \triangleq \frac{\mathbb{E}[N(k)]}{Ml}. \tag{13}$$

Let p^b be the popularity-priority policy and p^a be the coverage-priority policy. For each path $m \in \{1, \ldots, M\}$, we draw $Z_m \sim \text{Bernoulli}(\lambda)$ independently; if $Z_m = 0$ we sample the path using p^b, and if $Z_m = 1$ we sample it using p^a. Let ν^{p^b} and ν^{p^a} be the induced visitation distributions defined by Eq. 13 when all M paths are sampled exclusively from p^b and p^a, respectively. By linearity of expectation, the hybrid sampling induces the convex combination

$$\nu_{hybrid} = (1 - \lambda)\nu^{p^b} + \lambda \nu^{p^a}. \tag{14}$$

If the two base policies are constructed to induce the two base distributions,

$$\nu^{p^b} = \nu^{pop}, \qquad \nu^{p^a} = \nu^{cov}, \tag{15}$$

then Eq. 14 becomes

$$\nu_{hybrid}(k) = (1-\lambda)\nu^{pop}(k) + \lambda\nu^{cov}(k) = \nu^*(k), \tag{16}$$

which matches the unique optimizer of Eq. 9.

Remarks. A standard sufficient condition for Eq. 15 is that each policy corresponds to an ergodic Markov chain with stationary distribution ν^{pop} (resp. ν^{cov}), and each sampled path is initialized from the same stationary distribution. In this case, $\mathbb{P}(X_t = k) = \nu^{pop}(k)$ (resp. $\nu^{cov}(k)$) for every t, and the expected visitation distribution over any finite length l equals the stationary distribution exactly. More generally, if Eq. 15 holds approximately (e.g., finite-length paths with non-stationary initialization), then

$$\|\nu_{hybrid} - \nu^*\|_1 \leq (1-\lambda)\|\nu^{p^b} - \nu^{pop}\|_1 + \lambda\|\nu^{p^a} - \nu^{cov}\|_1, \tag{17}$$

so the deviation from the optimum is controlled by how well each base policy matches its target distribution.

Conclusion. Eq. 12 shows that the Euclidean-relaxed Knowledge Value trade-off objective admits a unique optimal distribution ν^* that is a convex combination of ν^{pop} and ν^{cov}. Eq. 14 shows that our path-level mixing induces, in expectation, a convex combination of the visitation distributions of the two base policies. Under Eq. 15, Algorithm 1 induces $\nu_{hybrid} \equiv \nu^*$, providing a rigorous theoretical justification for the effectiveness and interpretability of the parameter λ.

C Synthesis Model Selection Analysis

To systematically investigate the impact of different synthesis models, we conduct comparative experiments on strong open-source models. The open-source nature of these models enables large-scale deployment for synthesis, such as DeepSeek-R1 and Qwen2.5-72B-Instruct. We uniformly sample 1,000 seed instances across all difficulty levels and use both models to synthesize data for analysis.

DeepSeek-R1 achieves higher semantic diversity and greater diffusion from seed data. We first fix both models to synthesize $m = 10$ instances per seed (i.e., the same seed set and the same diffusion budget). Using Sentence-T5, we embed synthesized instances and compute the mean pairwise cosine distance over the entire synthesized set:

$$D_{\text{pairwise}} = \frac{1}{\binom{M}{2}} \sum_{1 \leq i < j \leq M} \left(1 - \cos(e_i, e_j)\right), \tag{18}$$

where $M = S \cdot m$ is the total number of synthesized instances for each model, and e_i denotes the embedding of the i-th synthesized instance. DeepSeek-R1 achieves $D_{\text{pairwise}} = 0.1732$, which is **1.24**$\times$ higher than Qwen's 0.1402.

Furthermore, we measure how far each model diffuses from an individual seed by the intra-seed spread. For each seed s, let $\{g_{s,i}\}_{i=1}^{m}$ be the embeddings of its m synthesized instances. We compute:

$$S_{\text{intra}}^{(s)} = \frac{1}{\binom{m}{2}} \sum_{1 \leq i < j \leq m} \left(1 - \cos(g_{s,i}, g_{s,j})\right), \tag{19}$$

and report the average over seeds, $S_{\text{intra}} = \frac{1}{S}\sum_{s=1}^{S} S_{\text{intra}}^{(s)}$. As illustrated in Figure A5, DeepSeek-R1 achieves an average intra-seed spread of 0.1175, which is **2.81**$\times$ higher than Qwen's 0.0418. Moreover, DeepSeek-R1 exhibits higher intra-seed spread than Qwen on **92.6%** of all seeds.

DeepSeek-R1 provides better coverage of seed knowledge points while introducing more novel ones. We annotate knowledge points (KPs) for both synthesized and seed data using the same KP labeling pipeline. Relative to the seed KP set, DeepSeek-R1 covers **23.4%** of seed KPs while introducing **842.2%** additional new KPs. In contrast, Qwen only covers 17.6% of seed KPs and introduces 553.6% new KPs.

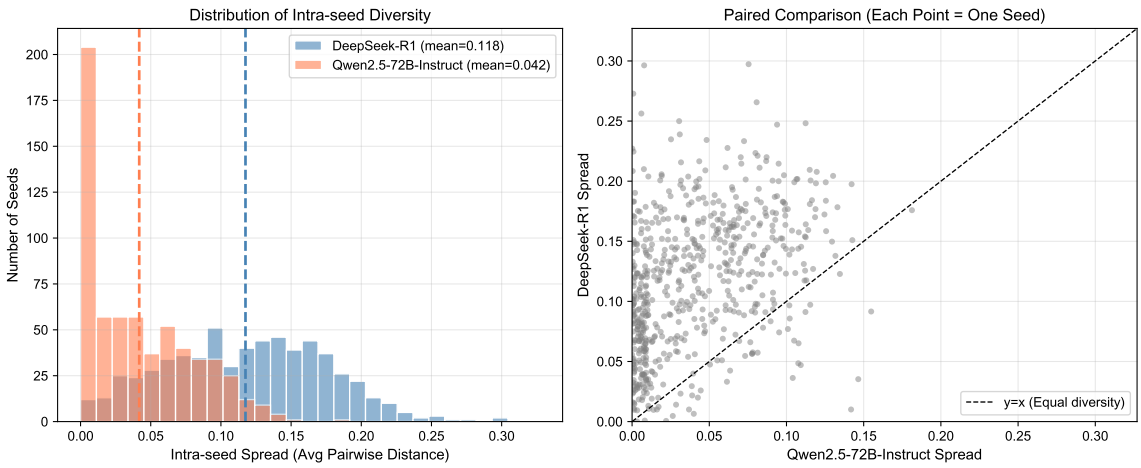

Figure A5: Embedding visualization of synthesized data. DeepSeek-R1 demonstrates greater semantic spread both globally and within individual seeds.

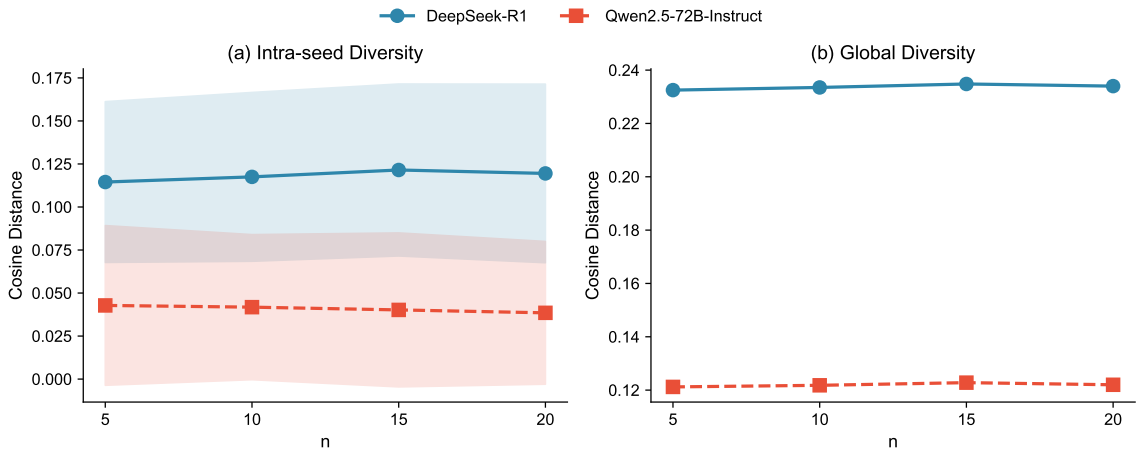

Figure A6: Diversity comparison between DeepSeek-R1 and Qwen across different diffusion numbers n. (a) Intra-seed diversity with standard deviation bands. (b) Global diversity. DeepSeek-R1 remains stable while Qwen's intra-seed diversity degrades as n increases.

DeepSeek-R1 maintains diversity as the diffusion number n increases, while Qwen degrades. We vary the per-seed diffusion number $n \in \{5, 15, 20\}$ and measure both intra-seed diversity (average pairwise cosine distance within each seed's outputs) and global diversity (average pairwise cosine distance across all outputs). As shown in Figure A6, DeepSeek-R1 maintains stable diversity metrics across different values of n, while Qwen's intra-seed diversity decreases by 10% (from 0.043 to 0.039) as n grows. Moreover, the gap between the two models widens with larger n: the R1/Qwen ratio increases from 2.68× to 3.10× for intra-seed diversity. This suggests that DeepSeek-R1 is more suitable for large-scale data synthesis without diversity degradation.

In summary, DeepSeek-R1 demonstrates clear advantages in synthesis diversity and knowledge coverage over Qwen2.5-72B-Instruct under controlled conditions. Therefore, we select DeepSeek-R1 as our data synthesis model for constructing LinkQA.

D Prompts

D.1 Data Annotation

To efficiently annotate the discipline labels while maintaining quality, we implement a two-stage annotation pipeline using the discipline classifier that categorizes content into 62 disciplines[2]. Initially, we employ DeepSeek-R1 with discipline-constrained prompts to generate preliminary labels for 20M seed samples. Subsequently, we curate a balanced subset of 500K high-confidence samples through uniform stratified sampling across all 62 disciplines, which is then used to finetune Qwen2.5-7B-Instruct, yielding our specialized subject classifier. Empirical validation demonstrates 82.18% label consistency between our specialized classifier and DeepSeek-R1, confirming reliable knowledge distillation. The prompt used to annotate data with discipline and train the corresponding labeler is shown as follows.

Prompt for Discipline Classifier

```
Act as an educational taxonomist. Classify the input question into our standardized discipline
 hierarchy using sequential reasoning, then output strictly in JSON format:
1. Primary Discipline Identification
   Select exactly one primary discipline from:
   {Discipline List}
   - Use "cross-discipline" only for explicit multi-domain integration
   - Assign "Other" only if no discipline matches >=60% relevance
2. Secondary Discipline Assignment
   - Identify the most specific applicable sub-discipline
   - Null if primary discipline has no sub-domains
   - Use "General" for non-specialized content
3. Validation Rules
   - Reject non-educational content -> Output "Invalid"
   - Correct spelling/terminology variations before classification
Output Schema:
{
  "primary_discipline": "",
  "secondary_discipline": "",
  "confidence": 0.0-1.0,
  "rejection_reason": null
}
Input: {Seed Data}
```

The list of 62 primary disciplines is as follows:

Discipline List (62)

```
['Mathematics', 'Computer Science and Technology', 'Clinical Medicine', 'Chemistry', '
Economics', 'Information Science and Systems Science', 'Physics', 'Biology', 'Law', '
Philosophy', 'Sociology', 'Literature', 'Psychology', 'Statistics', 'History', 'Power and
Electrical Engineering', 'Earth Science', 'Management Science', 'Electronics and Communication
 Technology', 'Linguistics', 'Preventive Medicine and Public Health', 'Political Science', '
Education Science', 'Aerospace Science and Technology', 'Astronomy', 'Materials Science', '
Mechanics', 'Sports Science', 'Ethnology and Cultural Studies', 'Basic Medicine', '
Environmental Science and Resource Science', 'Journalism and Communication', 'Religious
Studies', 'Engineering and Technology Related to Information and Systems Science', 'Food
Science and Technology', 'Engineering and Technology', 'Art Studies', 'Mechanical Engineering
', 'Traditional Chinese Medicine and Chinese Materia Medica', 'Pharmacy', 'Civil and
Architectural Engineering', 'Chemical Engineering', 'Nuclear Science and Technology', 'Marxism
', 'Agronomy', 'Energy Science and Technology', 'Transportation Engineering', 'Military
Science', 'Safety Science and Technology', 'Animal Husbandry and Veterinary Science', '
Archaeology', 'Engineering and Technology Related to Product Applications', 'Library,
Information and Documentation Science', 'Geomatics Science and Technology', 'Aquaculture
Science', 'Metallurgical Engineering Technology', 'Hydraulic Engineering', 'Military Medicine
and Special Medicine', 'Textile Science and Technology', 'Mining Engineering Technology', '
Forestry', 'Engineering and Technology Related to Natural Sciences']
```

[2]GB/T 13745-2008 taxonomy

The difficulty scorer operationalizes human performance metrics by defining five difficulty tiers (H1-H5) based on pass rates under standardized one-hour testing conditions with QS Top 100 university students majoring in relevant disciplines. We implement a multi-stage annotation pipeline where initial difficulty annotations are generated by DeepSeek-R1 through structured prompts that simulate human problem-solving behaviors, producing preliminary difficulty estimates for 500K QA pairs. Subsequently, we distill this knowledge by training Qwen2.5-14B-Instruct on the annotated data to create our specialized difficulty classifier. Expert assessment validation with five PhD evaluators (Krippendorff's $\lambda = 0.85$) confirms strong correlation (Pearson's $r = 0.92$) between model predictions and actual human performance metrics. The prompt used to annotate data with difficulty labels and train the corresponding labeler is shown as follows.

Prompt for Difficulty Scorer

```
Act as an educational assessment expert, analyze the provided question through sequential
reasoning and output strictly in JSON format:
1. Knowledge Analysis
   - Core concepts (<=3): [comma-separated list]
   - Integration type: {single-concept | cross-chapter | cross-discipline}
2. Cognitive Tier (Bloom's Taxonomy)
   {memory | understanding | application | analysis | synthesis | evaluation}
3. Difficulty Assessment
   - Estimated pass rate (P) for QS Top 100 university majors: [0-100%]
   - Tier:
     - extreme: P < 10%
     - challenge: 10% <= P < 30%
     - improvement: 30% <= P < 50%
     - standard: 50% <= P < 80%
     - basic: P >= 80%
     - other: invalid inputs
4. Exception Handling
   - Mark "other" for non-questions/unanswerable items
   - Correct minor errors (e.g., missing correct options) before assessment
   - Ignore provided solutions/answers

Output Schema:
{
  "difficulty_tier": "basic|standard|improvement|challenge|extreme|other",
  "rationale": [
    "Involves {N} core knowledge points",
    "Cognitive level: {Bloom's tier}",
    "Estimated pass rate: approximately {XX}% for target cohort"
  ]
}
Input: {Seed Data}
```

The knowledge point annotator identifies core knowledge concepts within educational content by extracting the most basic and smallest content units that constitute a knowledge system within specific discipline areas. We implement a multi-stage annotation pipeline where DeepSeek-R1 initially generates knowledge point labels for 20M seed samples using structured prompts that employ a hierarchical reasoning approach: first determining discipline classification and educational level assessment, then leveraging this contextual information to more accurately identify up to three core knowledge points per item. The annotation process employs educational taxonomist reasoning that analyzes content while ignoring potentially incorrect solutions, ensuring focus on authentic knowledge concepts. Subsequently, we curate a balanced training dataset through stratified sampling across multiple disciplines from the annotated seed data. This curated dataset is then used to finetune Qwen2.5-14B-Instruct, yielding our specialized knowledge point classifier that achieves an 80.08% agreement rate with DeepSeek-R1's annotations when allowing an edit distance of up to 3 on held-out test data. The prompt used to annotate data with knowledge point labels and train the corresponding classifier is shown as follows.

Prompt for Knowledge Point Annotation

```
Act as an educational taxonomist. Analyze the provided item through step-by-step reasoning and
 output strictly in JSON format:
1. discipline Classification
   - Identify the discipline to which the item belongs.
   - discipline list: {Discipline List}
2. Educational Level
   - Choose from: [Elementary School, Middle School, High School, University, Graduate School]
3. Knowledge Point Analysis
    - Core knowledge points (<=3): [comma-separated list]
    - Knowledge Point Definition: A knowledge point refers to the most basic and smallest
    content unit that constitutes a knowledge system within a certain discipline area.
    - Example:
     - Mathematics: Properties of linear functions
     - English: Present perfect tense
     - Biology: Basic laws of heredity
4. Exception Handling
   - Ignore any provided solutions or answer steps, as they may be incorrect or suboptimal.
   - Only select from the provided candidate lists for discipline, Assessment Ability, and
   Educational Level.
Output Schema:
{
  "Knowledge Point List": [
    "Properties of linear functions"
    ...
  ]
}
Input: {Seed Data}
```

D.2 Synthesis Prompts

The prompts and specific rules used to synthesize diverse knowledge-intensive QA pairs are as follows.

Prompt for Synthesizer

```
Act as a {Role Assigner} educator, analyze the knowledge points assessed by the provided {
ref_num} reference questions. Generate {gen_num} novel questions adhering to these
requirements:
1. Questions must demonstrate substantial differentiation while testing application or higher-
order use of identified knowledge points.
2. Difficulty must align with high-difficulty standards through:
   a) Down-scaling overqualified knowledge points to prerequisite concepts at graduate level
   b) Up-scaling underqualified points to advanced applications at graduate level
3. Linguistic consistency must be maintained with the input questions.
[Difficulty Reference Guide]
1. Knowledge Analysis:
   - Core concepts (<=3)
   - Integration type: {single | cross-chapter | cross-discipline}
2. Cognitive Tier (Bloom's Taxonomy):
   {memory | understanding | application | analysis | synthesis | evaluation}
3. Difficulty Calibration:
   - Estimate pass rate 0 <= P <= 100%
   - Tier Classification:
     - extreme: P < 10%
     - challenge: 10% <= P < 30%
     - improvement: 30% <= P < 50%
     - standard: 50% <= P < 80%
     - basic: P >= 80%
   - ENSURE generated questions match reference difficulty tier

Output Schema: {Format-specified JSON}
Input: {Seed Data}
```

{Role Assigner} can be set to "college" or "graduate". In our experiment, {ref_num} represents the number of reference QA pairs, and {gen_num} represents the number of synthetic QA pairs to generate. The typical mapping is: when {ref_num} = 1, {gen_num} = 10; when {ref_num} = 2, {gen_num} = 15;

when {ref_num} = 3, {gen_num} = 20.

Here is the prompt for regenerating answers to the generated questions:

Prompt for Answer Regenerator

```
Please strictly follow the requirements below to analyze the given question and answer:
Answer Requirements
1. Perform step-by-step reasoning and show the complete thought process, which must include:
   - Extraction of key information from the question
   - Application of relevant formulas/theorems
   - Analysis of each option individually
   - Reminders of common error types
   - Display of logical reasoning chains
2. Answer format requirements:
   - Must include both 'Solution Steps' and 'Final Answer' fields
3. Notes
   - If the question already includes solution steps and answers, please ignore them and don't
    be influenced by them, as they may be incorrect or suboptimal.
   - For multiple-choice questions:
     * If the correct answer is missing:
         - Add a fifth option: "(E) [Correct Answer]"
         - Set answer_index=4
         - Keep the original options unchanged
{Format-specific Constraints}
Output Schema: {Format-specified JSON}
Input: {Question}
```

{Format-specific Constraints} and {Format-specific JSON} are controlled by the rule enforcer and vary depending on whether the targeted synthetic question type is multiple-choice or essay-question format, following the specific rules below:

Rules for Targeted Synthetic Qustion Type

```
Format-specific Constraints:
Multiple-Choice: 4. The generated question type is multiple-choice. For each question, four
alternative options must be generated, and among the four options, there must be one correct
answer.
Essay-question: 4. The generated question type is essay-question. For each question, the
solution steps and the final correct answer are provided. The generated questions cannot be
open-ended questions (such as those of the solution type, thinking type, information listing
type, etc.), but must be self-contained with a final answer that can be determined as correct.

Format-specified JSON:
Multiple-Choice: [{"question": "", "options": [], "answer_index": 0-3}, ...]
Essay-Question: [{"question": "", "solution": "", "answer": ""}, ...]
```

E Case Study

Table A4: Case of different number of seed data and different question types.

Diffusion with one seed data (multiple-choice, n = 10)
Knowledge point path
Differential diagnosis of functional vs. organic abdominal pain
Seed Data 1
Question: I am a healthy 24-year-old female with a lifelong history of IBS. When I stand for a long time, I sometimes feel discomfort or heaviness under my left rib, in the upper abdomen, like there's a brick in there. It's not painful, just uncomfortable, especially after walking or standing for a couple of hours. What could be causing this and what should I do?

Table A4: (continued)

Answer: This discomfort could be related to your IBS, possibly due to gas accumulation or splenic flexure syndrome (trapped gas in the colon bend). Prolonged standing may worsen digestive symptoms or cause muscle strain. Consider: 1) Avoiding trigger foods and managing IBS symptoms, 2) Taking breaks during long periods of standing, 3) Gentle stretching and posture improvement. Consult your doctor if symptoms worsen or new symptoms develop, as proper evaluation may be needed to rule out other causes.

Synthetic QA pairs

Question 1: A patient with IBS complains of dull pain in the left upper abdomen after prolonged standing, with no peritoneal irritation on examination. What is the most likely underlying mechanism? A. Superior mesenteric artery compression syndrome B. Diaphragmatic reflux caused by increased intra-abdominal pressure C. Visceral hypersensitivity-induced referred pain D. Intercostal nerve entrapment syndrome

Answer 1: C. Visceral hypersensitivity-induced referred pain

Question 2: What is the key indicator to distinguish IBS-related abdominal pain from early chronic pancreatitis? A. Serum amylase fluctuation B. Temporal relationship between pain and defecation C. Abdominal CT findings D. Change in body mass index

Answer 2: B. Temporal relationship between pain and defecation

Question 3: When standing induces abdominal pain, which test is most specific for the diagnosis of hiatal hernia? A. 24-hour esophageal pH monitoring B. Standing abdominal ultrasound C. Enhanced CT 3D reconstruction D. Dynamic barium meal observation

Answer 3: D. Dynamic barium meal observation

Question 4: Which neural coordination disorder is most likely to cause postural (standing-induced) abdominal pain? A. Vagus nerve and greater splanchnic nerve B. Intercostal nerve and iliohypogastric nerve C. Pelvic nerve and pudendal nerve D. Phrenic nerve and subcostal nerve

Answer 4: A. Vagus nerve and greater splanchnic nerve

Question 5: What hemodynamic parameter should be especially noted when postural change induces abdominal pain? A. Portal venous pressure gradient B. Mesenteric vascular resistance C. Abdominal perfusion pressure D. Central venous pressure

Answer 5: B. Mesenteric vascular resistance

Question 6: For a patient with postural abdominal pain, which sign is key to distinguishing myogenic pain? A. Aggravation by Valsalva maneuver B. Positive rebound tenderness C. Induced by postural change D. Local muscle group tenderness

Answer 6: D. Local muscle group tenderness

...

Diffusion with one seed data (essay-question, n = 10)

Knowledge point path

Detection and dynamics of mini-moons around Earth

Seed Data

Question: Is there a large number of mini-moons around Earth? Due to being too small and too dim, and at distances of over 100,000 km from Earth, could many mini-moons remain undiscovered?

Answer: Currently, there is no conclusive evidence for a large number of natural mini-moons around Earth. While small objects like meter-scale asteroids can be temporarily captured by Earth's gravity (such as 2006 RH120, a 2-3 meter asteroid), these are typically transient and short-lived. No stable, naturally formed mini-moons exist in confirmed long-term orbits. The detection challenge you mention is valid - small, dim objects at great distances are difficult to observe. However, most captured objects follow unstable trajectories and eventually escape Earth's gravitational influence. As observation technology advances, we may discover more temporary companions, but a large population of stable natural mini-moons remains unlikely based on current orbital dynamics understanding.

Synthetic QA pairs

Question 1: If a ground-based telescope has a limiting magnitude of +25, and a spherical object with albedo 0.1 is at a distance of 150,000 km from Earth, what is the minimum detectable diameter using the law of reflection (ignore atmospheric extinction, Sun-Earth distance 1 AU, visible wavelength $\lambda = 550$ nm)?

Answer 1: The minimum detectable diameter is about 2.4 centimeters.

Question 2: Use dimensional analysis to derive the timescale τ for transient capture of mini-bodies as a function of Sun-Earth-Moon system parameters, including gravitational constant G, Earth's mass M, and lunar orbital radius a_m.

Table A4: (continued)

Answer 2: $\tau \propto \sqrt{a_m^3/(GM)}$, typically weeks to months.

Question 3: For a tidally-locked binary asteroid whose components are treated as homogeneous fluid ellipsoids, let the primary mass be M, the secondary mass be m, the orbital eccentricity be e, and the mass parameter be $\mu = m/(M+m)$. Using the Darwin–Radau equilibrium theory, derive the approximate expression for the maximum tidal (equatorial) elongation parameter δ in terms of e and μ.

Answer 3: The maximum tidal elongation parameter is $\delta \approx \frac{15}{4}\frac{\mu}{1+\mu}\, e$.

Question 4: A single two-way laser-ranging station in low-Earth orbit (altitude negligible compared to target distance) is used to detect 10 cm mini-moons at a geocentric range of 1 000 km. What is the required laser-pulse width (FWHM) for 10 cm single-shot range precision? Ignore angular resolution and assume only time-of-flight accuracy matters.

Answer 4: Required pulse width $\tau \leq 0.67$ ns.

Question 5: Compare the advantages of Delaunay variables and Poincaré variables in perturbation analysis of temporary satellite orbits, and state whether their Jacobian matrices are the identity.

Answer 5: Delaunay variables are convenient for Hamiltonian expansions but singular at $e = 0$ or $i = 0$; Poincaré variables remove both singularities via canonical transformation. Neither Jacobian is the identity matrix; both are symplectic.

Question 6: A spherical mini-moon of radius R and bulk density ρ orbits the Sun on a circular path of radius a. Derive the secular rate of change of its semi-major axis produced by direct solar radiation pressure of strength $P_\odot = L_\odot/(4\pi a^2 c)$, including the Bond albedo β.

Answer 6: $\dfrac{da}{dt} = -\dfrac{3(1+\beta)P_\odot}{2\rho R n}$.

...

Diffusion with two seed data (multiple-choice, n = 15)

Knowledge point path

Construction of contrapositive statements → De Morgan's laws

Seed Data 1	**Seed Data 2**
Question 1: Which of the following statements are correct? A. If p is a sufficient but not necessary condition for q, then $\neg p$ is a necessary but not sufficient condition for $\neg q$. B. Let $x, y \in \mathbb{R}$. The negation of the statement "If $xy = 0$, then $x^2 + y^2 = 0$" is true. C. The contrapositive of the statement "If $xy = 0$, then $x = 0$ and $y = 0$" is "If $x \neq 0$ and $y \neq 0$, then $xy \neq 0$". D. "$m = 1$ or $m = 2$" is the necessary and sufficient condition for the lines $(m+2)x + 3my + 1 = 0$ and $(m-2)x + (m+2)y - 3 = 0$ to be perpendicular.	**Question 2:** What is the contrapositive of a statement involving an OR condition? For example, what is the contrapositive of the statement: "For all dogs A, B, and C I have, if A and B are male, then B or C are Shibas"?
Answer 1: The correct statements are B and C.	**Answer 2:** The contrapositive of "If P, then Q" is "If $\neg Q$, then $\neg P$". Use De Morgan's law: $\neg(X \text{ or } Y) = (\neg X)$ and $(\neg Y)$.

Synthetic QA pairs

Question 1: What is the contrapositive of the statement 'If a number is prime greater than 2, then it is odd'? A. If a number is not odd, then it is not prime greater than 2. B. If a number is not prime greater than 2, then it is not odd. C. If a number is odd, then it is prime greater than 2. D. If a number is prime greater than 2, then it is not even.

Answer 1: A. If a number is not odd, then it is not prime greater than 2.

Question 2: According to De Morgan's laws, which expression is equivalent to $\neg(A \vee B)$? A. $A \wedge B$ B. $\neg A \vee \neg B$ C. $\neg A \wedge \neg B$ D. $A \vee B$

Answer 2: C. $\neg A \wedge \neg B$

Question 3: For the quantified statement '$\forall x$ (Student(x) → Studies(x))', what is the contrapositive? A. $\forall x(\neg\text{Student}(x) \rightarrow \neg\text{Studies}(x))$ B. $\forall x(\neg\text{Studies}(x) \rightarrow \neg\text{Student}(x))$ C. $\forall x(\text{Studies}(x) \rightarrow \text{Student}(x))$ D. $\forall x(\text{Student}(x) \wedge \neg\text{Studies}(x))$

Answer 3: B. $\forall x(\neg\text{Studies}(x) \rightarrow \neg\text{Student}(x))$

Table A4: (continued)

Question 4: If 'A or B' implies C, what is the contrapositive? A. If not C, then not A and not B. B. If not C, then not A or not B. C. If C, then A or B. D. If not A and not B, then not C.

Answer 4:A. If not C, then not A and not B

Question 5: Using De Morgan's laws, $\neg(P \land Q \land R)$ is equivalent to: A. $P \land Q \land R$ B. $\neg P \land \neg Q \land \neg R$ C. $P \lor Q \lor R$ D. $\neg P \lor \neg Q \lor \neg R$

Answer 5:D. $\neg P \lor \neg Q \lor \neg R$

Question 6: The contrapositive of 'If it is sunny or it is warm, I go outside' is: A. If I do not go outside, then it is not sunny and not warm. B. If I do not go outside, then it is not sunny or not warm. C. If I go outside, then it is sunny or warm. D. If it is not sunny and not warm, then I do not go outside.

Answer 6: A. If I do not go outside, then it is not sunny and not warm.

...

Diffusion with two seed data (essay-question, n = 15)

Knowledge point path

Procedures for submitting interim proposals → Rules for board meeting resolutions

Seed Data 1	**Seed Data 2**
Question 1: Which of the following statements about the organizational structure of a joint stock company is incorrect? A. When the company's unmade-up losses reach one third of the total paid-in capital, an interim shareholders' meeting shall be convened within two months. B. For a joint stock company to convene a shareholders' meeting, the time, place, and matters to be considered at the meeting shall be notified to all shareholders 20 days before the meeting. C. Shareholders who individually or collectively hold more than 3% of the company's shares may submit interim proposals in writing to the board of directors 10 days before the shareholders' meeting. D. A board meeting can only be held with more than half of the directors present, and resolutions must be passed by more than half of the attending directors.	**Question 2**: What are the requirements for holding a valid board meeting and passing a resolution in a joint stock company? A board meeting can only be held with more than half of the directors present, and resolutions must be passed by more than half of all directors.
Answer 1: A. According to corporate law, when a company's unmade-up losses reach one third of the total paid-in capital, an interim shareholders' meeting shall be convened within two months.	**Answer 2**: The statement is partially incorrect. A board meeting requires more than half of the directors present (correct), but resolutions must be passed by more than half of the **attending directors**, not more than half of all directors.

Synthetic QA pairs

Question 1: What is the minimum number of supervisors required in the board of supervisors of a joint stock company?

Answer 1: 3 supervisors

Question 2: What is the minimum percentage of shares that a single or collective shareholder of a joint stock company must hold to request the convening of an interim shareholders' meeting?

Answer 2: 3%.

Question 3: In a joint stock company, how many directors at least must be present for a board meeting to be held?

Answer 3: More than half of the directors

Question 4: What is the proportion of votes required for a general resolution to be passed at a shareholders' meeting in a joint stock company?

Answer 4: More than half of the voting rights of shareholders present

Table A4: (continued)

Question 5: When the company's unmade-up losses reach one third of the total paid-in capital, within how many days must an interim shareholders' meeting be convened?

Answer 5: Within 60 days.

Question 6: In a board meeting of a joint stock company with 11 directors and 7 present, how many directors must at least agree to pass a resolution?

Answer 6: 6 directors.

...

Diffusion with three seed data (multiple-choice, n = 20)

Knowledge point path

Poetry emotion comprehension → Imagery analysis → Theme analysis

Seed Data 1	Seed Data 2	Seed Data 3
Question 1: Read the following Song dynasty poem and answer: Which of the following appreciations are NOT appropriate? A. The first line uses 'crown and robe' to refer to the emperor, who is surrounded by ministers and stands above the people. B. The word 'only' in the second line depicts the ministers' cowardice. C. The couplet only praises Yongshu for his fairness, but also criticizes his lack of reason and rashness. D. The ending expresses the poet's best wishes for Yongshu. E. The poem is precisely constructed, deep in language, integrating narration, emotion, description, and commentary.	**Question 2**: Read the following Tang poem 'Wind and Rain' by Li Shangyin. What is the main emotion expressed in the poem?	**Question 3**: Read the following poem 'Apricot Blossoms by the North Slope' by Wang Anshi. What is the key imagery and philosophical meaning in the line 'Even if blown into snow by the spring wind'?
Answer 1: C, D	**Answer 2**: Melancholy, loneliness, and the hardships of life in exile.	**Answer 3**: The image of apricot blossoms as snow reflects a realm of subjective struggle and philosophical transcendence.

Synthetic QA pairs

Question 1: When comparing 'Guo po shan he zai' from Du Fu's 'Spring View' and 'Huang ye reng feng yu' from Li Shangyin's 'Wind and Rain', what is the common core technique and emotional effect? A. Both use virtual-real combination to express loneliness B. Both use contrast to highlight social injustice C. Both use natural imagery to symbolize national decline and strengthen patriotic feelings D. Both use allusions to express frustration at unrecognized talent

Answer 1: C. Both use natural imagery to symbolize national decline and strengthen patriotic feelings.

Question 2: According to Wang Guowei's 'theory of realms', which realm is best reflected in Wang Anshi's 'Even if blown into snow by the spring wind' in 'Apricot Blossoms by the North Slope'? A. Realm of no-self, emphasizing the unity of subject and object B. Realm of self, highlighting subjective struggle C. Writing the scene, focusing on realistic depiction D. Creating a realm, constructing an ideal world

Answer 2: A. Realm of no-self, emphasizing the unity of subject and object.

Question 3: The fundamental reason for the difference in opening techniques between Su Shi's 'The River Flows East' in 'Reminiscing at Red Cliff' and Xin Qiji's 'Eternal Joy at Jingkou North Pavilion' 'Thousand-year-old rivers and mountains' is? A. Su Shi emphasizes the passage of time, Xin Qiji emphasizes vastness of space B. Su Shi uses virtual description, Xin Qiji uses realistic description C. Su Shi expresses personal detachment, Xin Qiji emphasizes historical responsibility D. Su Shi is influenced by Taoism, Xin Qiji by Confucianism

Answer 3: C. Su Shi expresses personal detachment, Xin Qiji emphasizes historical responsibility

Question 4: Both Li Bai's 'Born to be useful' in 'Bring in the Wine' and Li Shangyin's 'Heartbroken at the wine of Xinfeng' in 'Wind and Rain' mention wine, but the difference in their attitudes towards life lies in? A. Li Bai is optimistic and enterprising, Li Shangyin is pessimistic and depressed B. Li Bai writes about the grandeur of the High Tang, Li Shangyin about the decline of the Late Tang C. Li Bai uses wine to express heroism, Li Shangyin uses wine to express sorrow D. Li Bai focuses on personal enjoyment, Li Shangyin focuses on social criticism

Answer 4: C. Li Bai uses wine to express heroism, Li Shangyin uses wine to express sorrow

Table A4: (continued)

Question 5: The artistic conception created by Wang Wei's 'After new rain in the empty mountains' in 'Autumn Evening in the Mountains' is similar to Tao Yuanming's 'Picking chrysanthemums by the eastern fence' in 'Drinking Wine', but the key philosophical difference is? A. Wang Wei blends Zen emptiness, Tao Yuanming embodies Confucian engagement B. Wang Wei emphasizes natural beauty, Tao Yuanming highlights the joy of seclusion C. Wang Wei is influenced by Buddhism, Tao Yuanming by Daoist non-action D. Wang Wei writes realistically, Tao Yuanming writes idealistically

Answer 5: C. Wang Wei is influenced by Buddhism, Tao Yuanming by Daoist non-action

Question 6: In Bai Juyi's 'The Song of the Lute Player', how is the emotional effect of 'We are all people fallen to the ends of the earth' mainly enhanced? A. By contrasting the fates of the musician and the poet B. By simulating life's ups and downs through musical rhythm C. By using the imagery of the ends of the earth to symbolize a wandering fate D. By direct expression to strengthen resonance

Answer 6: A. By contrasting the fates of the musician and the poet

...

Diffusion with three seed data (essay-question, n = 20)

Knowledge point path

Manifestations of racial segregation policy in the military → Colonial troops in World War I → Wartime social conditions and soldier adaptation

Seed Data 1	Seed Data 2	Seed Data 3
Question 1: What organization was Doris "Dorie" Miller a prominent figure in and what was his significant action during a critical event in that organization's history?	**Question 2**: Just another relapse post. (Long narrative about PTSD, trauma, and relapse during isolation, with emphasis on lack of support and coping mechanisms.)	**Question 3**: What role did colonial troops play in World War I and how did wartime social conditions affect soldier adaptation?
Answer 1: Doris "Dorie" Miller was a prominent figure in the United States Navy during World War II. He is best known for his heroic actions during the attack on Pearl Harbor, where he manned an anti-aircraft gun without prior training and assisted wounded sailors, becoming a symbol of courage and challenging racial stereotypes.	**Answer 2**: The narrative details a PTSD relapse triggered by isolation, referencing trauma, institutional abuse, and the importance of support systems and coping strategies.	**Answer 3**: Colonial troops played significant roles in World War I, often facing racial discrimination and challenging social conditions, which impacted their adaptation and psychological well-being.

Synthetic QA pairs

Question 1: A study on PTSD relapse rates during COVID-19 isolation found that 120 out of 200 participants with prior trauma history relapsed into self-harm. Calculate the odds ratio for relapse if 70 out of 150 participants without isolation triggers relapsed. Interpret the clinical significance.

Answer 1: Odds ratio = 1.714; Isolation triggers significantly increase relapse risk.

Question 2: Using DSM-5 criteria, diagnose PTSD for a veteran who experiences avoidance of war reminders, hypervigilance, and negative alterations in cognition for 3 months after a combat event. One symptom from each cluster must be specified.

Answer 2: No, PTSD diagnosis not met due to missing intrusion symptoms.

Question 3: In a historical analysis, if 30% of African American sailors during WWII were restricted to menial roles due to segregation policies, and Doris Miller's action increased public support for desegregation by 40% based on media archives, calculate the probability that a random policy change within 5 years was influenced by Miller. Assume base desegregation support was 20%.

Answer 3: Estimated probability = 40%.

Question 4: A cognitive-behavioral therapy (CBT) trial for PTSD had a 75% success rate in a sample of 80 patients. Calculate the 95% confidence interval for the true success rate and determine if it is statistically different from a control group rate of 60% (n=80). Use z-test for proportions.

Answer 4: 95% CI = (0.655, 0.845); Yes, significantly different from control.

Question 5: Based on the UN Convention Against Torture (CAT) definition—'severe pain or suffering, physical or mental, intentionally inflicted by or with the consent of a public official for purposes such as punishment, coercion or discrimination'—which scenario meets the criteria: (a) Solitary confinement in a hospital for 48 hours with no medical justification and under color of law, or (b) Verbal abuse by a family member?

Answer 5: Scenario (a) meets the UN torture criteria; (b) lacks state involvement and severity required under CAT.

Table A4: (continued)

Question 6: If the relapse rate for self-harm is 40% in PTSD patients without intervention, and CBT reduces it by 25% relative risk, calculate the absolute risk reduction (ARR) and number needed to treat (NNT) for CBT.

Answer 6: ARR = 10%; NNT = 10.

...

Table A5: Case of different difficulty of generate data.

H1

Sample 1: Question: An antibacterial drug achieves therapeutic target value when AUC/MIC=125. Given MIC=2mg/L, CL=10L/h, what is the total daily dose (F=1)?

Answer: Total daily dose is 2500mg

Sample 2: Question: When merging two sorted singly linked lists into one sorted list, what is the minimum number of pointer reassignments required in the worst-case scenario?

A. O(1) B. O(log n) C. O(n) D. O(n log n)

Answer: C. O(n)

H2

Sample 1: Question: A company's current stock price is $80, with expected earnings per share of $5.00 next year and a dividend payout ratio of 40%. If the market required rate of return is 10%, calculate the implied perpetual growth rate. If the actual growth rate is 5%, determine if the stock price is reasonable.

Answer: The implied growth rate is 7.5%, and the current stock price is overvalued.

Sample 2: Question: If using game theory to model the collapse of the feudal system, which equilibrium best describes the struggle for power among lords?

A. Nash equilibrium (individual optimality leading to collective suboptimality) B. Pareto optimality (no improvement without harming others) C. Coordination game (cooperation driven by common interests) D. Prisoner's dilemma (betrayal as dominant strategy)

Answer: D. Prisoner's dilemma (betrayal as dominant strategy)

H3

Sample 1: Question: Design a machine learning-based algorithm to predict skin retraction rate after liposuction. Which biomechanical parameters should be selected? Explain the parameter selection criteria and algorithm architecture.

Answer: Key parameters include biomechanical factors (initial dermal strain ε_0, stress relaxation time τ, elastic modulus E, skin thickness), physiological factors (age, BMI, collagen density, skin hydration), and surgical parameters (liposuction volume, anatomical location). Parameter selection criteria: clinical relevance, biomechanical significance, and measurability. Architecture: data preprocessing $\rightarrow$ feature engineering $\rightarrow$ ensemble model (XGBoost + Random Forest) $\rightarrow$ output layer (retraction rate %) $\rightarrow$ model interpretation layer. The model incorporates 15+ parameters with cross-validation and SHAP analysis for clinical interpretability.

Sample 2: Question: Given two coprime integers m and n, what is the minimal k such that there exist unique pairs (x,y) in {1,2,...,k} satisfying x $\equiv$ a mod m and y $\equiv$ b mod n?

A. m + n B. m + n - 1 C. m * n D. lcm(m,n)

Answer: B. m + n - 1

H4

Sample 1: Question: A child with congenital hypothyroidism still shows delayed intellectual development after treatment. Analyze possible reasons from a molecular mechanism perspective.

Answer: Possible mechanisms: 1: MCT8 transporter defects causing T3 deficiency in brain tissue; 2: THRB mutations causing hormone resistance; 3: Treatment initiated after the critical period of 2 weeks after birth.

Sample 2: Question: Matrix A = [[1, 2], [0, 1]] transforms the curve $x^2 + y^2 = 1$ into a new curve C. What is the minimum distance between points on C and the line x - y = 3?

A. (3 - $\sqrt{2}$)/$\sqrt{2}$ B. $\sqrt{2}$ C. 3/$\sqrt{2}$ D. 3 - $\sqrt{2}$

Answer: A. (3 - $\sqrt{2}$)/$\sqrt{2}$

H5

Table A5: (continued)

Sample 1: Question: Formulate the hydrodynamic limit of the Boltzmann equation to the incompressible Navier-Stokes equations using the Chapman-Enskog expansion. Derive the viscosity and heat conductivity coefficients in terms of the collision kernel.

Answer: Viscosity $\mu = \frac{1}{15}\int_{\mathbb{R}^3}\int_{S^2}|v-v_*|^\gamma b(\cos\theta)\sin^2\theta d\theta dv$ and heat conductivity $\kappa = \frac{5}{2}\mu$ for Maxwell molecules ($\gamma = 1$), following the Chapman-Enskog procedure.

Sample 2: Question: When constructing the contraction $\phi_{|A|} : \overline{M}_{0,n}(X,\beta) \to Y_{|A|}$, if A is a Kawamata divisor class on $\overline{M}_{0,n+m}/\mathfrak{S}_m$, what is the necessary condition for the Picard rank of $Y_{|A|}$ to be 1?

A. β is an extremal curve class B. Picard rank of X is 1 C. $\mathfrak{S}_m$ acts freely D. All boundary divisors $\Delta_{i,j}$ are contracted $(i,j \neq 2, n+m-2)$

Answer: D. All boundary divisors $\Delta_{i,j}$ are contracted $(i,j \neq 2, n+m-2)$

Table A6: Samples for synthetic QA pairs of different disciplines. For each discipline, Sample 1 is multiple-choice and Sample 2 is essay- question.

Mathematics

Sample 1: Question: An e-commerce platform has discount rules: 50 yuan off when spending 299 yuan, 100 yuan off when spending 499 yuan, with coupons stackable. For an order with items totaling 630 yuan and shipping fee of 15 yuan, what is the minimum payment amount after applying discounts?

Answer: 545

Sample 2: Question: Which principle ensures that the outer measure of the set of random reals is 1 in both V and V[G], preventing it from being 0?

A. Baire Category Theorem B. Kolmogorov's Zero-One Law C. Fubini's Theorem D. Borel Determinacy

Answer: A. Baire Category Theorem

Computer Science and Technology

Sample 1: Question: In 3D stacked chips, analyze the impact of Through-Silicon Via (TSV) on clock signal integrity, and propose three methods to suppress reflection noise with their comparative advantages and disadvantages.

Answer: Termination resistors (stable but area-consuming), stepped TSV (efficient but difficult to fabricate), pre-emphasis (flexible but requires additional circuitry).

Sample 2: Question: How to solve the spatial audio mismatch problem caused by audio-video separation when processing VR 360° videos?

A. Add Ambisonic metadata B. Force mono audio output C. Enable head-related transfer function D. Increase spatial audio delay compensation

Answer: A. Add Ambisonic metadata

Clinical Medicine

Sample 1: Question: For a newborn with bilateral persistent eye discharge without redness or fever, how should anatomical characteristics of the lacrimal duct and microbiological features be integrated for differential diagnosis?

Answer: Analysis must consider the unopened Hasner's valve at the distal nasolacrimal duct (causing sterile mucus accumulation) and bacterial colonization risks (such as C. trachomatis vertical infection incubation characteristics), using cytological examination of secretions (neutrophil predominance suggests infection) and pathogen culture for differentiation.

Sample 2: Question: When comparing the response differences to LASIK surgery between children and adult amblyopia patients, which of the following is the most fundamental reason for poorer prognosis in adults?

A. Decreased corneal healing capacity B. Closure of neural plasticity window C. Lower refractive error stability D. Reduced visual system redundancy

Answer: B. Closure of neural plasticity window

Chemistry

Sample 1: Question: Calculate the Gibbs free energy change (ΔG) during zymogen activation, given that peptide bond hydrolysis releases -5 kJ/mol and conformational reorganization requires +3 kJ/mol. Determine if the process is spontaneous and explain its physiological significance.

Answer: $\Delta G = -2$ kJ/mol < 0, so the process is spontaneous. Physiological significance: The hydrolysis reaction drives conformational changes, ensuring activation occurs only under specific triggering conditions, preventing accidental activation.

Table A6: (continued)

***Sample 2**:* Question: Given the battery reaction: Zn + $2H^+ \rightarrow Zn^{2+}$ + H_2, with standard electromotive force of 0.76 V. If conducted in a buffer solution at pH=5, with P_{H_2}=1 atm and [Zn^{2+}]=0.1 M, the actual electromotive force is approximately: (Given $E^{\circ}_{Zn^{2+}/Zn}$ = -0.76 V)

A. 0.50 V B. 0.64 V C. 0.76 V D. 0.88 V

Answer: A. 0.50 V

Economics

***Sample 1**:* Question: After a country implements a universal basic income policy, how can one distinguish between the direct economic effects and indirect health awareness effects of this policy on the increased use of preventive medical services? Please propose a regression model specification that includes instrumental variables.

Answer: Use a 2SLS model with pre-implementation regional characteristics as instrumental variables, treating total income change as an endogenous variable. By comparing the significance of economic variable coefficients with health awareness variable coefficients, decompose direct economic effects and indirect awareness effects.

***Sample 2**:* Question: In an environment of rising interest rate expectations, a company plans to buy back stocks and hold floating-rate debt. What is the most effective derivative strategy to hedge the overall risk?

A. Buy Interest Rate Cap B. Sell Interest Rate Swap (Pay Fixed, Receive Floating) C. Buy Treasury futures D. Sell Credit Default Swap (CDS)

Answer: A. Buy Interest Rate Cap

Information Science and Systems Science

***Sample 1**:* Question: For metal parts in additive manufacturing, how to build a digital twin system for quality prediction based on acoustic emission signals? Explain the feature extraction algorithm and real-time simulation architecture.

Answer: Key features: 1 : Energy ratio in the 100-150kHz frequency band 2 : Hilbert-Huang marginal spectral entropy value 3 : Principal components from singular value decomposition of joint time-frequency distribution.

***Sample 2**:* Question: In a negative feedback system with open-loop transfer function $G(s) = \frac{K}{s(s+2)}$, where $K > 0$, what condition must K satisfy for the closed-loop system to be stable?

A. $K < 2$ B. $K > 2$ C. $K < 4$ D. $K > 4$

Answer: D. $K > 4$

Physics

***Sample 1**:* Question: In the relativistic framework, when an object moves at 0.8c, what is the ratio of its relativistic mass to rest mass? If a force perpendicular to the direction of motion is applied at this time, derive the expression for acceleration.

Answer: Relativistic mass ratio is 5/3; transverse acceleration $a_{\perp}$ = $3F/(5m_0)$

***Sample 2**:* Question: Gravity is described as spacetime curvature, while electromagnetic force is transmitted through photons. What is the key contradiction when unifying these two forces at the quantum scale? Which higher-order theoretical framework needs to be introduced?

A. Contradiction in force range; string theory B. Contradiction in relativistic effects; quantum field theory C. Contradiction in energy conservation; grand unified theory D. Contradiction in time asymmetry; loop quantum gravity

Answer: A. Contradiction in force range; string theory

Biology

***Sample 1**:* Question: Explain why a protein solution that has been heat-sterilized might regain activity after cooling, while a protein solution treated with pepsin cannot recover its activity.

Answer: Heat denaturation doesn't break covalent bonds, allowing potential renaturation upon cooling; pepsin digestion hydrolyzes peptide bonds causing irreversible structural damage.

***Sample 2**:* Question: Comparing the developmental cycles of Chlamydia and Rickettsia, what is the key difference in their reproductive morphology energy metabolism? Integrate biochemical pathway reasoning.

A. Chlamydial elementary bodies depend on host ATP, while Rickettsia perform autonomous oxidative phosphorylation B. Both utilize glycolysis for energy, but Chlamydia at higher rates C. Rickettsia reticulate bodies have mitochondria-like structures, while Chlamydia lack them D. Chlamydia utilize photosynthesis, while Rickettsia depend on amino acid degradation

Answer: A. Chlamydial elementary bodies depend on host ATP, while Rickettsia perform autonomous oxidative phosphorylation

Law

Table A6: (continued)

***Sample 1**:* Question: According to relevant provisions of the Company Law and Securities Law, if Company A controls 60% of Company B's voting rights through an agreement, and Company B holds 15% shares of listed Company C. When Company A increases its shareholding in Company C to 8%, is it required to make a tender offer? Please analyze based on the rules for identifying persons acting in concert.

Answer: Triggered. A and B constitute persons acting in concert, with combined shareholding of 23% (15%+8%). Although below 30%, after the increase, every additional 5% requires suspension and announcement. The trap in this question is the cumulative calculation rule and the regulation of shareholding increases.

***Sample 2**:* Question: Limited partner Qian engaged in a competitive business and profited, but used the partnership's trade secrets. General partner Sun claimed the profits belonged to the partnership. If the partnership agreement does not stipulate this situation, according to the Anti-Unfair Competition Law and Partnership Enterprise Law, which of the following is correct?

A. All profits belong to the partnership due to infringement B. Part of the profits belong to the partnership, with the proportion determined by court C. Qian bears no responsibility as competitive business is implicitly allowed D. Sun must prove intentional infringement to seek compensation

Answer: B. Part of the profits belong to the partnership, with the proportion determined by court

Table A7: Samples for synthetic QA pairs of different knowledge points. For each discipline, Sample 1 is multiple-choice and Sample 2 is essay- question.

Single Knowledge Point

***Sample 1**:*

Knowledge Point: Legal Reservation Principle

Question: An art gallery requires teachers in group visits to present an introduction letter from their institution, while students only need student IDs. Some teachers sued for infringement of equal rights. Please analyze the legality of the introduction letter requirement based on the principle of legal reservation.

Answer: Illegal. The requirement for an introduction letter lacks legal basis and constitutes an undue restriction on teachers' rights, violating the principle of administrative legality.

***Sample 2**:*

Knowledge Point: Noether Normalization Lemma

Question: Noether normalization lemma states that a finitely generated k-algebra contains a polynomial subalgebra over which it is integral. In the geometric interpretation, what does this imply about the corresponding variety?

A. It is unirational B. It is affine and irreducible C. It has a finite morphism to affine space D. It is smooth in codimension one

Answer: C. It has a finite morphism to affine space

Two Knowledge Points

***Sample 1**:*

Knowledge Points: Acid-Base Balance Disorders, Water-Electrolyte Balance Disorders

Question: A patient with respiratory failure developed altered consciousness after mechanical ventilation. Emergency tests showed: blood sodium 128mmol/L, chloride 88mmol/L, arterial blood gas pH 7.52, $PaCO_2$ 28mmHg, HCO_3^- 24mmol/L. Diagnose the acid-base imbalance type and explain the mechanism of ionic abnormalities.

Answer: Acute respiratory alkalosis; low sodium and chloride suggest concurrent dilutional hyponatremia, possibly due to inappropriate antidiuretic hormone secretion after ventilation improvement.

***Sample 2**:*

Knowledge Points: AIC Model Selection Criteria, Linear Regression Model Assumptions

Question: When comparing nested linear models (Model A: basic features, Model B: adds polynomial terms) using AIC, which outcome indicates that the polynomial terms are justified, and what is a key assumption?

A. AIC decreases by more than 2, assuming errors are normally distributed and independent. B. AIC increases, but p-values for polynomial terms are significant, indicating overfitting. C. R-squared improves marginally, requiring BIC to confirm model parsimony. D. F-statistic is insignificant, implying polynomial terms add no value despite AIC change.

Answer: A. AIC decreases by more than 2, assuming errors are normally distributed and independent.

Three Knowledge Points

Table A7: (continued)

***Sample 1**:*

Knowledge Points: Phase Field Method Fundamentals, Anisotropic Phase Field Models, Free Energy Function Construction

Question: Within the phase field method framework, derive the control equations for a polycrystalline growth model that generates Voronoi-like structures. With grain boundary energy anisotropy coefficient < 0.05, provide the phase field variable evolution equation and free energy function form.

Answer: The control equation is the anisotropic Allen-Cahn equation, with a free energy function containing gradient terms, double-well potential, and intergranular repulsion terms. Anisotropy is modulated through $\varepsilon(\theta)$.

***Sample 2**:*

Knowledge Points: Isolating Switch Operation Risk Analysis, Power System Transient Process, Surge Arrester and Voltage Transformer Configuration Principles

Question: In a 220kV double-busbar connection, if surge arresters and voltage transformers share one set of isolating switches, the maximum operational risk occurs when:

A. The isolating switch operates under load and produces an arc B. Voltage fluctuation caused by surge arrester operation during lightning strikes C. Short circuit on the secondary side of voltage transformer D. Transient process during busbar switching

Answer: A. The isolating switch operates under load and produces an arc