

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

SCHEMA FOR IN-CONTEXT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

In-Context Learning (ICL) enables transformer-based language models to adapt to new tasks by conditioning on demonstration examples. However, traditional example-driven in-context learning lacks explicit modules for knowledge retrieval and transfer at the abstraction level. Inspired by cognitive science, specifically schema theory, which holds that humans interpret new information by activating pre-existing mental frameworks (schemas) to structure understanding, we introduce SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-ICL). This proposed framework extracts the representation of the Building Blocks of Cognition for the reasoning process instilled from prior examples, creating an abstracted schema — a lightweight, structured template of key inferential steps and their relationships — which is then used to augment a model’s reasoning process when presented with a novel question. We demonstrate that a broad range of large language models (LLMs) lack the capacity to form and utilize internal schema-based learning representations implicitly, but instead benefit significantly from explicit schema-based scaffolding. Across chemistry and physics questions from GPQA dataset, our empirical experiment results show that SA-ICL consistently boosts performance (up to 39.67%) when the single demonstration example is of high quality, which simultaneously reduces reliance on the number of demonstrations and enhances interpretability. SCHEMA-ACTIVATED IN-CONTEXT LEARNING not only bridges disparate ICL strategies ranging from pattern priming to Chain-of-Thought (CoT) prompting, but also paves a new path for enhancing human-like reasoning in LLMs.

1 INTRODUCTION

In-Context Learning (ICL) has emerged as a dominant approach for adapting large language models (LLMs) to new tasks without requiring fine-tuning or additional parameter updates. By conditioning on a set of demonstrations, ICL enables LLMs to leverage prior knowledge and generalize to unseen examples. Despite its effectiveness, traditional ICL does not align fully with how humans acquire and apply knowledge in real-world learning scenarios, as it lacks mechanisms for episodic memory and context-rich encoding (Li et al., 2024a).

Learning in humans is inherently structured, involving knowledge abstraction, retrieval, and adaptive reasoning. Research in cognitive science suggests that humans develop mental frameworks, called schemas, that organize prior knowledge and facilitate problem solving in new contexts (Rumelhart and Ortony, 1977). These schemas enable efficient retrieval of relevant information and guide interpretation and action, reducing reliance on explicit demonstrations (Rumelhart and Ortony, 1977). Critically, schema activation, consisting of bringing the proper schema into working memory, is essential for effective comprehension and analogical transfer; retrieval alone may not suffice Gick and Holyoak (1983); Gentner (1983).

Recent evidence in the behavior of LLMs mirrors this limitation. For instance, recent models such as GPT-4 retrieve numerous plausible analogs with high recall, but often select incorrect ones due to their reliance on surface-level similarity rather than structural alignment (Puranam et al., 2025). This setback calls for mechanisms that go beyond retrieval, mobilizing schema-like abstractions to guide reasoning.

Inspired by these cognitive insights, we introduce SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-ICL), a schema-driven retrieval and reasoning framework for language models. Rather than

054 retrieving task-specific demonstrations alone, SA-ICL first guides the model to construct a problem
 055 schema, which acts as its corresponding structured abstraction. Prior examples are then retrieved
 056 using these schemas as keys, and critically, their schemas are then used to activate and refine the
 057 schema of the new problem. This activation process enables LLMs to solve problems more ef-
 058 ficiently by integrating structured prior knowledge into current reasoning, addressing the structural
 059 mapping gap observed in analogical reasoning of LLMs (Puranam et al., 2025). Echoing approaches
 060 using latent graph schemas for fast transfer learning (Guntupalli et al., 2023), our method enforces
 061 structured abstraction as the medium for retrieval, reasoning, and inference.

062 A particularly important domain for this approach is scientific reasoning. Although physics and
 063 chemistry may appear distinct, their foundational problem-solving strategies often converge on
 064 shared relational structures. A prime example is the existence of a “conservation law” schema, a
 065 structural template for identifying initial and final states around a core principle. Using schema
 066 activation, SA-ICL enables cross-domain transfer, where schemas developed in one scientific field
 067 can scaffold reasoning in another, similar to analogical transfer, which depends on mapping hidden
 068 relational structures rather than surface similarities (Kang et al., 2025).

069 We evaluate SA-ICL on the graduate-level scientific benchmark *Graduate-Level Google-Proof*
 070 *Q&A* (GPQA dataset) (Rein et al., 2024), which consists of challenging PhD-level physics, chem-
 071 istry, and biology multiple-choice questions that require structured reasoning. Our experiments are
 072 particularly focused on the physics and chemistry subsets of GPQA dataset. The results show that
 073 leveraging SA-ICL enhances accuracy compared to standard ICL (One-Shot) in most scenarios.
 074 Most notably, our framework improves accuracy by up to 39.67% over One-Shot for chemistry ques-
 075 tions and by up to 34.45% for physics questions, when the retrieved examples are of high similarity.
 076 Importantly, we demonstrate that the One-Shot prompting alone does not provide the optimal gain
 077 in LLMs’ performance during in-context learning, whereas utilizing activated schemas consistently
 078 improves reasoning efficiency and effectiveness, especially when the knowledge density is high. We
 079 further analyze model outputs to illustrate the interpretability benefits of schema activation.
 080

Our contributions are as follows:

- 081 • We propose SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-ICL), a novel frame-
 082 work that integrates schema construction, schema-guided retrieval, and schema activation
 083 to enable more efficient, generalizable, and interpretable in-context learning.
- 084 • We perform comprehensive evaluations comparing SA-ICL with traditional ICL and
 085 retrieval-only baselines across multiple scientific reasoning benchmarks, demonstrating
 086 consistent improvements in overall accuracy.
- 087 • We provide interpretability analyses showing that schema activation facilitates more struc-
 088 tured reasoning processes and reduces token reliance, offering a path toward more efficient
 089 inference-time reasoning.

090 Overall, our findings suggest that SA-ICL advances beyond example-driven ICL by bridging re-
 091 trieval with schema activation, mirroring human cognitive strategies more closely. By leveraging
 092 abstract schemas to refine reasoning rather than relying solely on explicit demonstrations, schema-
 093 driven ICL reduces dependence on examples and makes inference more efficient and interpretable.

095 2 RELATED WORK

097 2.1 SCHEMA THEORY IN COGNITIVE PSYCHOLOGY

099 Schema theory is one of the fundamental theories of cognitive psychology. This theory conceptual-
 100 izes how humans organize and structure knowledge into coherent mental frameworks, or schemas,
 101 that are constructed from prior experiences and serve as interpretive structures for understanding
 102 new information Rumelhart and Ortony (1977). These abstract structures are dynamic; they actively
 103 guide how prior knowledge is encoded and retrieved and how the new information is perceived
 104 (Brewer and Treyens, 1981). Classic research in psychology from decades ago, like Bartlett (1932)
 105 and Piaget (1952) established that human learning involves either interpreting new information into
 106 existing schemas (*assimilation*), or modifying these existing schemas to incorporate novel knowl-
 107 edge (*accommodation*). Activated schemas from the lens of prior knowledge provide a cognitive
 108 mechanism that enables efficient problem-solving and reasoning by guiding the retrieval process

108 and allowing individuals to make inferences and fill in missing details (Anderson and Pichert, 1978;
 109 Piaget, 1952). This model of human cognition, where activating the correct abstract structure is key
 110 to interpreting a new problem, provides the direct theoretical motivation for the SA-ICL framework.
 111

112 **2.2 IN-CONTEXT LEARNING METHODOLOGIES**

113 **2.2.1 EXAMPLE-DRIVEN APPROACHES (E-ICL)**

114 One-Shot and few-shot learning paradigms (Brown et al., 2020) have been adopted as computa-
 115 tionally efficient methodologies (Parnami and Lee, 2022) for enabling language models to perform
 116 inference-time in-context learning without requiring internal parameter updates. Example-driven
 117 ICL utilizes predefined question-answer pairs to change the probability distribution of output tokens
 118 conditioned on user queries and prior knowledge (Wang et al., 2020; Min et al., 2022b).
 119

120 Despite state-of-the-art (SOTA) LLMs achieving substantially extended context windows, few-shot
 121 learning continues to demand extensive computational resources when the number of demon-
 122 stration samples and their associated token counts increase, resulting in inevitable computational cost
 123 inflation (Keles et al., 2023). Furthermore, LLMs demonstrate sensitivity to performance worsen-
 124 ing when processing long in-context demonstrations for complex reasoning tasks, making example-
 125 driven ICL a compromise for tasks characterized by complex reasoning processes (Li et al., 2024b).
 126 While example-driven ICL establishes a connection between human-interpretable prompting and
 127 machine learning, LLMs require additional mechanisms to achieve full alignment between their
 128 computational processes and human cognitive patterns (Mahowald et al., 2024).
 129

130 The majority of existing ICL research — including MetaICL and PCW — conceptualizes LLMs pri-
 131 marily as pattern-matching systems operating over prompt examples, without comprehensive anal-
 132 ysis of their internal abstraction mechanisms (Min et al., 2022a; Ratner et al., 2023). Current ap-
 133 proaches show fundamental limitations, including high task specificity and response rigidity, rather
 134 than enabling a generalization across diverse domains or a naturalistic use of real-world knowledge
 135 (Yang et al., 2022). Additionally, empirical experiments indicate that traditional example-driven ICL
 136 achieves optimal performance only when context lengths extend to hundreds of thousands of tokens
 137 through multi-shot prompting (Agarwal et al., 2024). SA-ICL advances the exploration of how lan-
 138 guage models can generate reasoning processes through their internal knowledge representations in
 139 a human-interpretable manner, simultaneously optimizing for performance quality, computational
 140 cost, and token efficiency.

141 **2.2.2 ABSTRACTION-DRIVEN APPROACHES**

142 While most existing in-context learning methods were heavily example-driven, previous works
 143 raised key issues (Saglam et al., 2025; Lampinen et al., 2024; Dong et al., 2022). Recently, the
 144 machine learning (ML) community has been witnessing a growing development of in-context learn-
 145 ing approaches in a broader perspective (Lampinen et al., 2024), including abstraction-driven in-
 146 context learning (A-ICL) (Swaminathan et al., 2023), which could contribute to understanding the
 147 way models understand and utilize context. Although previous works have linked the mechanism
 148 of induction heads in LLMs to the contextual maintenance and retrieval (CMR) model in human
 149 episodic memory (Olsson et al., 2022; Polyn et al., 2009), direct evidence for high-level schema
 150 induction in language models remains limited.

151 In contrast, our work adopts an A-ICL approach that explicitly extracts general reasoning steps,
 152 conducting experiments at a higher level than task-specific knowledge. Prior studies have shown that
 153 fixed, structured generation — enabled by carefully designed prompting mechanisms or properly
 154 constrained decoding strategy — can improve LLMs’ performance on reasoning tasks (e.g., ReAct,
 155 Program-of-Thoughts) (Yao et al., 2023; Chen et al., 2023). Furthermore, retrieval of previous LLM-
 156 generated schemas or demonstrations has been shown to be autoregressively beneficial for ICL, as
 157 in prompt-retrieval and retrieval-augmented methods (Rubin et al., 2022; Shi et al., 2024).
 158

159 **2.2.3 CHAIN-OF-THOUGHT REASONING**

160 Chain-of-Thought (CoT) reasoning has recently arisen as a critical strategy within ICL by explicitly
 161 outlining intermediate steps before arriving at a final answer to enhance the reasoning capabilities
 (Wei et al., 2022; Kojima et al., 2022), thus significantly improving LLM performance on tasks that

162 require multi-step inference (Nye et al., 2021; Wang et al., 2023). However, CoT reasoning usually
 163 operates within example-driven reasoning frameworks, where explicit reasoning details are provided
 164 through a few demonstrations to guide model outputs (Zhang et al., 2023). Various methods have
 165 been explored in recent studies to optimize and extend CoT. Multiple thought paths are sampled for
 166 reasoning and aggregated before inference to improve output reliability in Self-Consistency Prompt-
 167 ing (Wang et al., 2023). Moreover, in Least-to-Most prompting, reasoning is progressively refined
 168 by starting with simpler sub-questions (Zhou et al., 2022).

169 Our proposed framework, SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-ICL), although
 170 sharing conceptual foundations, constitutes a fundamentally distinct computational paradigm. Tra-
 171 ditional CoT prompting operates as an instance-specific methodology wherein the model constructs
 172 novel and frequently verbose linear reasoning sequences tailored to each input query, consequently
 173 necessitating substantial human engineering of input queries to achieve satisfactory performance on
 174 specific tasks (Stechly et al., 2024). The instance-specific trajectories also exhibit limited gener-
 175 alizability across disparate task domains (Stechly et al., 2024). In contrast, SA-ICL incorporates
 176 structured schema retrieval mechanisms, enabling language models to access and retrieve abstract,
 177 generalized schemas from their prior knowledge, thereby automatically adapting reusable cognitive
 178 abstractions to novel problems during inference, rather than depending upon task-specific demon-
 179 strations. SA-ICL serves two purposes: enhancing interpretability and facilitating knowledge trans-
 180 fer. These characteristics render SA-ICL particularly effective for complex reasoning tasks requir-
 181 ing high-level conceptual abstraction, including scientific inquiry and hypothesis generation.
 182

183 2.3 OTHER HUMAN-INSPIRED PROMPTING METHODS

184 Wang and Zhao (2024) applied human introspective reasoning strategies by splitting the question-
 185 answer queries into multiple metacognitive prompting steps to improve LLMs' capability in ques-
 186 tion understanding. This work explores the problem of understanding the gap between human and
 187 LLM reasoning processes. Zhou et al. (2023) prompting attempted to address the knowledge loss
 188 for LLMs in tasks with chaotic input contexts, where relevant information is obscured by distrac-
 189 tors, by guiding LLMs to segment and analyze the input systematically, summarizing the findings
 190 as they go, before drawing an answer, to reduce the knowledge loss in long-context scenarios ef-
 191 fectively. Retrieval-Augmented Generation (RAG) provides LLMs with access to prior knowledge
 192 within a given knowledge base for future queries on similar tasks, which can be considered long-
 193 term memories for LLMs. However, traditional RAG limits the quality of the retrieval strategy and
 194 the corresponding reasoning logic learned from prior knowledge examples by the quality of ex-
 195 isting knowledge base examples. It remains a challenge for existing RAG techniques to adapt to
 196 dynamic and interconnected knowledge bases Gutiérrez et al. (2025). **SA-ICL** is built on top of
 197 the RAG paradigm and leverages schema theory, which humans use to adapt to the dynamic and
 198 interconnected knowledge base, by retrieving abstracted reasoning logic from memorized examples
 199 for activating a schema for the new problem. Our work emphasizes using human cognitive schemas
 200 to fill in the knowledge gap of LLMs between their perceived examples and similar tasks, where
 201 the ground truth answer is not apparent in the input context. In contrast, the knowledge needed is
 202 closely related.

203 3 SCHEMA-ACTIVATED IN-CONTEXT LEARNING

205 We propose this innovative ICL framework, SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-
 206 ICL), which mimics how humans use previous examples to activate a schema, enabling a better
 207 understanding and solution to a new problem. This framework can be applied to any trained large
 208 language model and combined with existing prompting techniques. This framework is simple, yet
 209 also flexible and extendable, providing a reliable and transparent explanation of how a language
 210 model learns from previous examples and turns these abstractions into a powerful schemas that
 211 guides it in solving new problems.

212 3.1 OVERALL WORKFLOW

213 SA-ICL operationalizes schema theory from cognitive science in five steps, aligning abstract
 214 schema formation with language model retrieval and reasoning. (i) **Problem Representation:** Given
 215

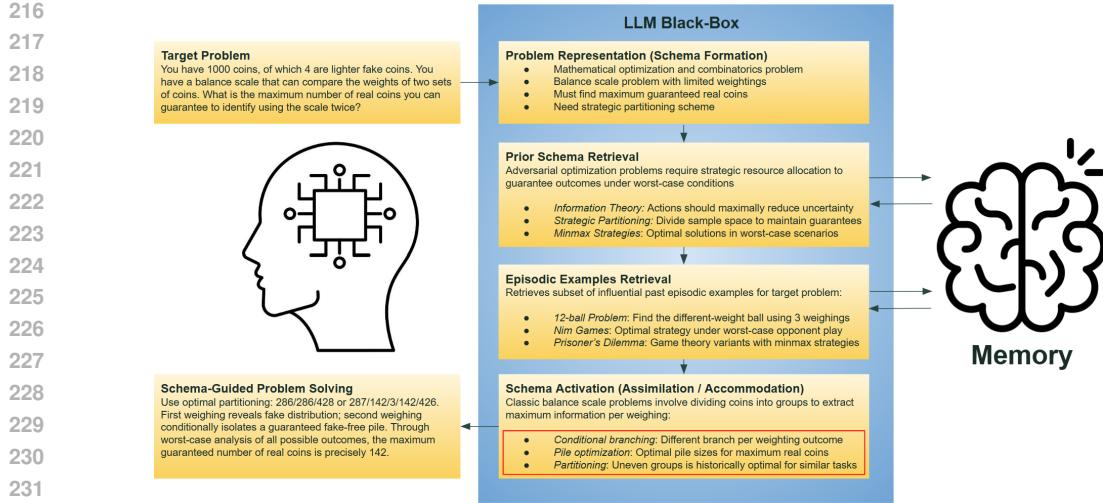


Figure 1: SA-ICL applied to a coin-weighing optimization problem. The framework demonstrates five sequential stages: (i) generates problem representation to form an initial schema recognizing this as a mathematical optimization and partitioning challenge, (ii) retrieves a similar prior schema to identify relevant properties of the question, (iii) gathers a subset of past influential episodic examples that are relevant to solve the target problem, (iv) performs schema activation and integrates retrieved knowledge to develop optimal mathematical strategies adapted to solve the target problem, and (v) conditions the target problem to the adapted schema and utilizes schema-guided inference to eventually conclude that the maximum number of guaranteed real coins is 142.

Algorithm 1 SCHEMA-ACTIVATED IN-CONTEXT LEARNING

Require: Problem x ; Schemas $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_N\}$; Episodic set $\mathcal{E} = \{e_1, \dots, e_M\}$; Memory $\mathcal{M} = \{(\mathcal{S}_i, e_j, w_{ij}(t))\}$; Representation \mathcal{R} ; Similarity sim; Activation f ; LLM; Threshold $\tau \in [0, 1]$

Ensure: $y, \mathcal{S}_{\text{new}}$

- 1: $\mathcal{S}_x \leftarrow \mathcal{R}(x)$
- 2: $\hat{\mathcal{S}} \leftarrow \mathcal{S}_x$
- 3: $\hat{\mathcal{E}}_\tau \leftarrow \emptyset$
- 4: **for** $j \in \{1, \dots, M\}$ **do**
- 5: **if** $w_{ij}(t) \geq \tau$ **then**
- 6: $\hat{\mathcal{E}}_\tau \leftarrow \hat{\mathcal{E}}_\tau \cup \{e_j\}$
- 7: **end if**
- 8: **end for**
- 9: **end for**
- 10: $\mathcal{S}_{\text{new}} \leftarrow f(\mathcal{S}_x, \hat{\mathcal{S}}, \hat{\mathcal{E}}_\tau)$
- 11: $y \leftarrow \text{LLM}(x, \mathcal{S}_{\text{new}})$
- 12: **return** $(y, \mathcal{S}_{\text{new}})$

an input problem x , the LLM constructs a representation $\mathcal{S}_x = \mathcal{R}(x)$. This representation functions as the initial schema for the new problem. (ii) **Prior Schema Retrieval:** SA-ICL retrieves the most relevant schema $\hat{\mathcal{S}} \in \mathcal{S}$ that maximizes similarity with \mathcal{S}_x . (iii) **Episodic Examples Retrieval:** Conditioned on the retrieved schema $\hat{\mathcal{S}}$, SA-ICL collects a subset of episodic examples whose decayed association weights $w_{ij}(t)$ exceed a threshold τ . This yields a set $\hat{\mathcal{E}}_\tau$ of examples that remain influential for the current reasoning. (iv) **Schema Activation (Assimilation / Accommodation):** The retrieved schema $\hat{\mathcal{S}}$ and episodic set $\hat{\mathcal{E}}_\tau$ are integrated with the current problem representation, producing a new activated schema: $\mathcal{S}_{\text{new}} = f(\mathcal{S}_x, \hat{\mathcal{S}}, \hat{\mathcal{E}}_\tau)$. This integration may proceed through *assimilation* when prior schema fits well, or *accommodation* when internal restructuring is required. (v) **Schema-Guided Problem Solving:** Finally, the LLM solves the task by conditioning on the input x and the adapted schema \mathcal{S}_{new} : $y = \text{LLM}(x, \mathcal{S}_{\text{new}})$.

Figure 1 describes the conceptual pipeline of SA-ICL. Section A details the complete mathematical formalization of each step in the framework. Algorithm 1 summarizes SA-ICL framework.

270 4 MAIN EXPERIMENTS
271272 4.1 TECHNICAL SET UP
273274 All experiments used standardized OpenAI-style API endpoints. For local runs, we used an NVIDIA
275 A40 GPU with 24GB RAM. In addition to model inference for QWen-3 and Llama-3.1, all em-
276 bedding generation and reranker computations were also performed on the A40 GPU, with results
277 cached locally for faster loading. Section B summarizes the model families we used and the corre-
278 sponding execution environments.279 4.2 TASKS AND METRICS
280281 **Tasks.** We designed closed-ended multiple-choice question-answering tasks wherein language mod-
282 els received individual questions per iteration and applied different reasoning approaches before
283 generating final answers. The experimental design incorporated multiple knowledge density levels
284 to evaluate model performance across varying degrees of prior knowledge acquisition. Closed-ended
285 questions were selected to ensure fair comparison between baseline methods and the SA-ICL.286 **Datasets.** The experiments primarily utilized GPQA dataset, a rigorously annotated benchmark con-
287 taining questions in chemistry and physics that were subjected to comprehensive human annotation.
288 The chemistry subset was employed for initial refinement of schema-based prompting strategies,
289 while the physics subset was also used during the experiment stage. These subsets are designated
290 as *GPQA-Chemistry* and *GPQA-Physics*, respectively. To simulate scenarios with dense knowledge
291 bases, GPT-4o was employed to generate three synthetic variants for each problem in the database
292 through criteria-based prompting (Section C). The criteria defined three distinct similarity levels:
293 *Essentially Same*, *Similar*, and *Different*. The synthetic datasets are designated *GPQA-Chemistry-
294 Synthetic* and *GPQA-Physics-Synthetic*. Section D showed the similarity between the problems in
295 the synthetic datasets and the target problems in the GPQA dataset.296 **Evaluation.** For each independent question, the final responses were compared with the ground
297 truths, which led to a downstream performance. We acknowledge that this is an indirect metric.
298299 4.3 EXPERIMENTAL SETUP
300

301 4.3.1 HIGH QUALITY EXAMPLES

302 The first specific experiment discussed in this paper was a direct response to our research ques-
303 tion: *Are examples all we need?* In particular, we investigated the LLMs' performances when the
304 examples were of high quality.
305306 We acknowledge the inherent challenges in curating high-quality exemplars for effective model
307 prompting. To address this, we adopted two distinct strategies:308 1. **Synthetic Similarity:** We generated synthetic data using controlled prompts to simulate
309 varying levels of similarity between the generated examples and the target questions (see
310 Section C for detailed prompt templates). We refer to this as *synthetic similarity*.
311 2. **Latent Similarity:** We used Cohere's Rerank 3.5 to retrieve semantically related exam-
312 ples from GPQA dataset and the synthetic pool, employing cross-encoder rerankers (see
313 Section D for detailed analysis). We refer to this as *latent similarity*.314 In this experiment, the LLMs were provided with *Essentially Same* questions as One-Shot examples.
315 We then compare this result with the LLM groups that were provided with schemas. We reported
316 the gaps to answer the question and argue that examples alone were not always sufficient.
317318 4.3.2 DENSITY OF KNOWLEDGE BASE ON SCHEMA & DOWNSTREAM PERFORMANCES
319320 To better understand the extent to which SA-ICL depends on the quality of examples when leverag-
321 ing One-Shot strategies, we conducted systematic experiments by varying the density of examples
322 provided to the LLMs. More specifically, the levels of example relevance from highest to lowest
323 quality for both approaches are listed below:324 • **Synthetic Similarity:** *Essentially Same* → *Similar* → *Different*

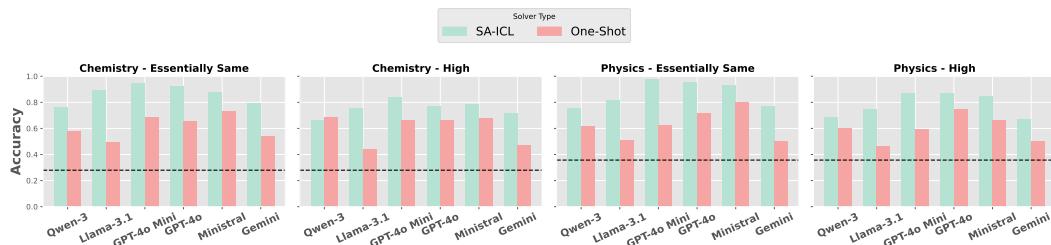


Figure 2: The performances of six LLMs on Chemistry and Physics questions using SA-ICL and example-driven One-Shot prompting, with two retrieval methods (Section 4.3.2). The black dotted line represents the zero-shot performance.

• Latent Similarity: *High* → *Medium* → *Low*

Notably, we found that a substantial portion of *High* examples correspond to *Essentially Same* (61.18% in chemistry and 67.44% in physics). This experiment investigates both whether SA-ICL is still useful even when the example is of low quality, and when SA-ICL results in the greatest benefits. In our experiments with Latent Similarity, *High* indicates there is only one example that is of the highest similarity, while *Medium* includes four other examples that had a smaller similarity scores and *Low* includes eight other examples that are less similar to the question compared to the top@1 example.

4.3.3 INTERPRETABILITY EXPERIMENTS

As an investigation into the underlying reasoning mechanisms in SA-ICL, we conducted interpretability experiments to analyze decision patterns in model reasoning and confidence in outputs. We selected a representative chemistry problem from *GPQA-Chemistry* and compared three in-context learning techniques: (1) One-Shot + schema, (2) pure One-Shot, and (3) One-Shot + CoT.

For each approach, we extracted token-level probability distributions over multiple-choice options from the model’s output logits. This analysis reveals the confidence each method has in arriving at its final answer and looks into whether SA-ICL achieves correct responses through schema activation or surface-level pattern matching. We also analyzed response length and structure to evaluate reasoning efficiency, testing our hypothesis that SA-ICL enables more direct knowledge access compared to the verbose reasoning typically required by CoT approaches.

5 MAIN RESULTS

5.1 HIGH-QUALITY EXAMPLES DO NOT LEAD TO OPTIMAL PERFORMANCE WITHOUT SCHEMA ACTIVATION

Language models achieved moderate accuracy when only provided with high-quality examples (Section 4.3.2), yet their performance remained suboptimal compared to models employing explicit schema-based learning techniques (Figure 2). This performance gap suggests that relying solely on models’ internal representations for high-level abstraction fails to fully exploit the learning potential of quality demonstrations. When models were conditioned to implement schema-based learning through structured guidelines and templates (Algorithm 1), we observed consistently significant improvements in downstream task performance. These performance gains were consistent even when using domain-agnostic schemas without task-specific fine-tuning for physics or chemistry (Sections E and F). We also conducted an experiment with GPT-5 on a different benchmark (Section G).

5.2 DENSITY OF KNOWLEDGE AFFECTS THE PERFORMANCE OF SA-ICL

The experimental results presented in Section 5.3 demonstrate a stratified performance profile for the SA-ICL methodology. Under optimal conditions, when the provided exemplar exhibits essential equivalence to the target question, the SA-ICL approach achieves complete dominance with a

378 perfect win rate against the One-Shot baseline. In sub-optimal conditions characterized by varying
 379 degrees of quality and similarity, the SA-ICL method maintains consistent performance advantages,
 380 demonstrating superior results across the majority of experimental conditions even when exemplar-
 381 question alignment is imperfect.

382 These findings indicate that while SA-ICL achieves maximum efficacy when provided with highly
 383 relevant exemplars, its performance benefits extend beyond ideal matching conditions. The evidence
 384 supports the conclusion that SA-ICL constitutes a fundamentally superior prompting methodology
 385 rather than a technique dependent solely on exceptional example quality, establishing its viability as
 386 a generalizable improvement to existing in-context learning frameworks.
 387

388 5.3 EXTENDING THE EXPERIMENTS TO DATASETS OF MEDXPERTQA, MMLU, AND 389 COMMONSENSEQA 390

391 To evaluate the effectiveness of SA-ICL in a broader domains, in particular when the knowledge
 392 base is dense. We have conducted evaluations with MedXpertQA (Zuo et al., 2025), MMLU (Wang
 393 et al., 2024), and CommonSenseQA (Talmor et al., 2019). We have seen a consistent improvement
 394 by SA-ICL as shown in Table 2. We also compared with One-Shot with Reasoning in the context
 395 and 5-Shots and observed that SA-ICL could be a more effectiveness context learning approach
 396 compared with the CoT Reasoning path and SA-ICL exploits one single best knowledge, reducing
 397 the necessity for have multiple-shots, in the dense knowledge scenerios.
 398

399 5.4 ABLATION STUDY: EFFECTIVENESS OF THE ACTIVATION 400

401 We conducted an ablation study to investigate whether it is the abstraction of the example or the
 402 activation of the schema that boosts LLMs’ performance. Since we observed the most significant
 403 margin in performance between SA-ICL and One-Shot with the GPT-4o mini, we removed the
 404 activation part. We only used the abstraction of the example to solve the Physics questions. We
 405 observed that LLMs would not perform as well as when they were allowed to activate the schema
 406 Figure I.4. This result highlights the importance of the explicit human-like schema activation.
 407

408 5.5 A DEEPER INVESTIGATION INTO THE EFFECTIVENESS OF SA-ICL 409

410 To investigate the interpretable effectiveness of SA-ICL over CoT and One-Shot prompting meth-
 411 ods, we logged the log likelihood of the top-5 predicted tokens in the LLM generation output (Fig-
 412 ure H.3)(Zhang et al., 2025). One-Shot prompting often encourages the model to overfit to the given
 413 demonstration by focusing on fitting the output format (e.g., Answer: ANSWER), rather than lever-
 414 aging the example to activate prior knowledge and understandings to reason the problem better and
 415 provide well-thought-out answers. This problem is discussed in earlier work as pattern-matching.
 416

417 While prior work noted that enforcing rigid, structured outputs can reduce the flexibility needed
 418 for effective reasoning, SA-ICL reached equilibrium between structured reasoning and free-form
 419 thinking. Table 3 showed that in high-knowledge settings, where the answer is already well repre-
 420 sented in the context, CoT may introduce unnecessary verbosity and even hurt performance, whereas
 421 SA-ICL provides more direct and efficient knowledge activation.
 422

423 6 DISCUSSION 424

425 SCHEMA-ACTIVATED IN-CONTEXT LEARNING (SA-ICL) challenges the conventional paradigm
 426 of machine learning, which has historically relied on large quantities of demonstrations. This coun-
 427 terintuitive shift has constrained the development of abstraction-driven approaches, even in the era of
 428 LLMs. However, from a cognitive perspective, SA-ICL demonstrates a stronger alignment with hu-
 429 man thinking patterns. Our empirical results demonstrate that activated schemas, viewed through the
 430 lens of prior knowledge, enhance the effectiveness of that knowledge when appropriately matched to
 431 the problem domain. We demonstrated that schemas enrich the contextual abstraction of examples,
 432 and this process facilitates LLMs’ understanding of schema generation mechanisms, which sub-
 433 sequently conditions these models to utilize schemas alongside examples more efficiently. In these
 434 scenarios, a single example augmented with schema activation yields performance improvements

432 Table 1: Performances of SA-ICL and One-Shot on Chemistry and Physics questions. For Chemistry
 433, the improvement is up to **39.67%**, **34.88%** for Physics. On average, the improvement in
 434 Chemistry is **9.81%**, **12.91%** for Physics. All values in the table were rounded up to the third deci-
 435 mal. Note that for Gemini, the One-Shot, in the Latent Similarity, we were Gemini 2.0 due to the
 436 Gemini 1.5 Flash being deprecated.

(a) Chemistry Results

Model	Method	Synthetic Similarity			Latent Similarity		
		Essentially Same	Similar	Different	High	Medium	Low
Qwen-3	SA-ICL	0.763	0.376	0.301	0.667	0.634	0.624
	One-Shot	0.581	0.301	0.301	0.688	0.581	0.624
Llama-3.1	SA-ICL	0.892	0.430	0.387	0.753	0.548	0.495
	One-Shot	0.495	0.366	0.366	0.441	0.441	0.473
GPT-4o Mini	SA-ICL	0.946	0.462	0.366	0.839	0.581	0.559
	One-Shot	0.688	0.366	0.323	0.667	0.624	0.613
GPT-4o	SA-ICL	0.925	0.516	0.419	0.774	0.581	0.667
	One-Shot	0.656	0.559	0.409	0.667	0.688	0.699
Minstral	SA-ICL	0.882	0.473	0.376	0.785	0.634	0.624
	One-Shot	0.731	0.376	0.280	0.677	0.656	0.656
Gemini 1.5 Flash*	SA-ICL	0.796	0.333	0.280	0.452	0.663	0.640
	One-Shot	0.538	0.258	0.194	0.473	0.452	0.409

(b) Physics Results

Model	Method	Synthetic Similarity			Latent Similarity		
		Essentially Same	Similar	Different	High	Medium	Low
Qwen-3	SA-ICL	0.756	0.465	0.349	0.686	0.814	0.721
	One-Shot	0.616	0.477	0.477	0.605	0.581	0.581
Llama-3.1	SA-ICL	0.814	0.430	0.407	0.744	0.535	0.605
	One-Shot	0.512	0.314	0.372	0.465	0.395	0.547
GPT-4o Mini	SA-ICL	0.977	0.512	0.523	0.872	0.581	0.628
	One-Shot	0.628	0.372	0.372	0.593	0.593	0.547
GPT-4o	SA-ICL	0.953	0.663	0.616	0.872	0.674	0.698
	One-Shot	0.721	0.616	0.547	0.744	0.698	0.721
Minstral	SA-ICL	0.930	0.535	0.488	0.849	0.686	0.624
	One-Shot	0.802	0.372	0.256	0.663	0.721	0.686
Gemini 1.5 Flash*	SA-ICL	0.767	0.407	0.360	0.655	0.559	0.559
	One-Shot	0.500	0.349	0.349	0.500	0.488	0.512

468
 469 exceeding 20% compared to using the example alone, indicating that SA-ICL substantially reduces
 470 the number of examples required for pattern matching compared to traditional ICL approaches.
 471

472 It is worth highlighting that although the improvements of SA-ICL are most significant when the
 473 episodic examples exhibit high similarity, our framework exhibits noticeable increases in accuracy
 474 compared to One-Shot in most scenarios overall. However, there are still special circumstances
 475 where accuracy boosts are not observed. We hypothesize this is because our current implementation
 476 generates schemas from single examples ($\tau = 1$), which may cause the model to apply reasoning
 477 patterns too rigidly without adequate contextual grounding. Nevertheless, through the deployment of
 478 the complete SA-ICL algorithm, we posit that dynamic schema activation mechanisms will enhance
 479 performance even when the knowledge space exhibits sparsity. We encourage the LLM community
 480 to pursue this direction toward developing models with more human-like cognitive capabilities.
 481

482 Finally, it should be noted that human perception of the environment around them typically does not
 483 begin with textual information as the initial sensory input that triggers their inherent schema-based
 484 thinking. Instead, visual and other sensory data captured through computer vision systems would
 485 likely raise even further the necessity for SA-ICL in real-world deployment scenarios.

486
 487 Table 2: Performances of SA-ICL, One-Shot, One-Shot with Reasoning and 5-Shots on MedX-
 488 pertQA, MMLU, and CommonSenseQA questions. Note that for the MMLU questions were the
 489 original MMLU questions included in the MMLU-Pro dataset, for MedXpertQA, we were using the
 490 Skeletal questions, for CommonSenseQA, we randomly picked up 200 questions from the validation
 491 subset.

492	Dataset	Method	Accuracy
493	MMLUCollegeMath	SA-ICL	0.7973
494		One-Shot	0.6216
495		One-Shot with Reasoning	0.7162
496		5-Shots	0.7027
497	MedXpertQA	SA-ICL	0.7342
498		One-Shot	0.6034
499		One-Shot with Reasoning	0.6667
500		5-Shots	0.7027
501	CommonSense	SA-ICL	0.9200
502		One-Shot	0.9000
503		One-Shot with Reasoning	0.8400
504		5-Shots	0.8800

506 Table 3: Token counts and correctness (**✓/✗**) across different prompting strategies for the first 10
 507 questions from *GPQA-Chemistry* dataset using GPT-4o Mini with temperature set to 0 for the most
 508 consistent results. All prior knowledge in this experiment is retrieved using *High* in Latent Similar-
 509 ity. We ran **three** experiments per question to get the average token counts, and we used the majority
 510 correctness as overall correctness.

512	Question ID	SA-ICL		One-Shot		One-Shot + CoT	
		Tokens	Correct	Tokens	Correct	Tokens	Correct
514	2662eff7a6231613f...caaeb	150	✓	133	✗	196	✗
515	fc081c2fb63be500...65420	161	✓	132	✓	206	✓
516	a8be7a4963fb6bc7...99122	180	✓	156	✓	228	✓
517	f730b35adb897658b...a77e5	166	✓	412	✗	417	✗
518	1ce3d847d25b2c2f6...01155	231	✓	235	✓	254	✓
519	d8c36bd55ba561cb4...7a049	308	✓	324	✓	273	✓
520	40b2b50a3c993902d...0bfcb	91	✓	138	✓	186	✓
521	a2136b05b78259562...184d7	76	✓	76	✗	91	✗
522	cbf5c336a0990294b...7d447	203	✓	170	✗	222	✗
	16464cac7090a24d3...9baf2	175	✓	154	✓	217	✓
523	Total Correct	10/10		6/10		6/10	

525 7 REPRODUCIBILITY STATEMENT

526 The full code will be posted on GitHub after the review is done. It is worth noting that Gemini
 527 1.5 will be deprecated on Sep 24, 2025. After this date, it will no longer be possible to reproduce
 528 the results from our experiments using Gemini 1.5 models. However, the authors will neverthe-
 529 less provide full experiment results with Gemini 1.5 to the public. To reproduce the experiments,
 530 you can follow the README.md under the zipped code submission in the supplementary mate-
 531 rial. The supplementary material also includes the raw experimental results and consists of multiple
 532 CSV files for the experiments in Section 4.3, which we used to analyze and report the results in
 533 Section 5.3. To examine the raw LLM responses, you can run the following Python command:
 534 `base64.urlsafe_b64decode(process_id).decode()`.
 535

540 REFERENCES
541

542 Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang,
543 Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in
544 Neural Information Processing Systems*, 37:76930–76966, 2024.

545 Richard C Anderson and James W Pichert. Recall of previously unrecallable information following
546 a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1):1–12, 1978.

547 Frederic C Bartlett. *Remembering: A study in experimental and social psychology*. 1932.

548 William F Brewer and James C Treyens. Role of schemata in memory for places. *Cognitive Psy-
549 chology*, 13(2):207–230, 1981.

550 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhari-
551 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agar-
552 wal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,
553 Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz
554 Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec
555 Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In
556 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neu-
557 ral Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.,
558 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

559 Wenhui Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompt-
560 ing: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on
561 Machine Learning Research*, 2023.

562 Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu,
563 Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*,
564 2022.

565 Dedre Gentner. Structure-mapping: A theoretical framework for analogy. In *Cognitive Science*,
566 1983.

567 Mary L. Gick and Keith J. Holyoak. Schema induction and analogical transfer. *Cognitive Psychol-
568 ogy*, 15(1):1–38, 1983.

569 J. Swaroop Guntupalli et al. Graph schemas as abstractions for transfer learning, inference, and
570 planning. *ArXiv preprint*, 2023.

571 Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory:
572 Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*,
573 2025.

574 Hyeonsu B Kang, David Chuan-En Lin, Yan-Ying Chen, Matthew K Hong, Nikolas Martelaro, and
575 Aniket Kittur. Biospark: Beyond analogical inspiration to llm-augmented transfer. In *Proceedings
576 of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2025.

577 Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational
578 complexity of self-attention. In *International conference on algorithmic learning theory*, pages
579 597–619. PMLR, 2023.

580 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
581 language models are zero-shot reasoners. *Advances in neural information processing systems*,
582 35:22199–22213, 2022.

583 Andrew Kyle Lampinen, Stephanie CY Chan, Aaditya K Singh, and Murray Shanahan. The broader
584 spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024.

585 Ji-An Li, Corey Zhou, Marcus Benna, and Marcelo G Mattar. Linking in-context learning in trans-
586 formers to human episodic memory. *Advances in neural information processing systems*, 37:
587 6180–6212, 2024a.

594 Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with
 595 long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024b.

596

597 Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and
 598 Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in*
 599 *cognitive sciences*, 2024.

600 Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn
 601 in context. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz,
 602 editors, *Proceedings of the 2022 Conference of the North American Chapter of the Associa-*
 603 *tion for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle,
 604 United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.
 605 naacl-main.201. <https://aclanthology.org/2022.naacl-main.201/>.

606 Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke
 607 Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In
 608 Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference*
 609 *on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United
 610 Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.
 611 emnlp-main.759. <https://aclanthology.org/2022.emnlp-main.759/>.

612 Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David
 613 Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work:
 614 Scratchpads for intermediate computation with language models. 2021.

615 Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan,
 616 Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli,
 617 Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane
 618 Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish,
 619 and Chris Olah. In-context learning and induction heads. *CoRR*, abs/2209.11895, 2022. <https://doi.org/10.48550/arXiv.2209.11895>.

620

621 Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-
 622 shot learning. *ArXiv*, abs/2203.04291, 2022. <https://api.semanticscholar.org/CorpusID:247318847>.

623

624 Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin
 625 Zhang, Mohamed Shaaaban, John Ling, Sean Shi, et al. Humanity’s last exam. *arXiv preprint*
 626 *arXiv:2501.14249*, 2025.

627

628 Jean Piaget. *The origins of intelligence in children*. International Universities Press, 1952.

629

630 Sean M Polyn, Kenneth A Norman, and Michael J Kahana. A context maintenance and retrieval
 631 (cmr) model of organizational processes in free recall. *Psychological Review*, 116(1):129–156,
 632 2009.

633 Phanish Puranam, Prothit Sen, and Maciej Workiewicz. Can llms help improve analogical reasoning
 634 for strategic decisions? *ArXiv preprint*, 2025.

635

636 Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas,
 637 Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large
 638 language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceed-*
 639 *ings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*
 640 *Long Papers*), pages 6383–6402, Toronto, Canada, July 2023. Association for Computational
 641 Linguistics. doi: 10.18653/v1/2023.acl-long.352. <https://aclanthology.org/2023.acl-long.352/>.

642

643 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Di-
 644 rani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a bench-
 645 mark. In *First Conference on Language Modeling*, 2024.

646

647 Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context
 648 learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Associa-*
 649 *tion for Computational Linguistics: Human Language Technologies*, pages 2655–2671, 2022.

648 David E Rumelhart and Andrew Ortony. Schemata: The building blocks of cognition. In Rand J
 649 Spiro, Bertram C Bruce, and William F Brewer, editors, *Theoretical issues in reading compre-*
 650 *hension*, pages 33–58. Lawrence Erlbaum Associates, 1977.

651

652 Baturay Saglam, Xinyang Hu, Zhuoran Yang, Dionysis Kalogerias, and Amin Karbasi. Learning
 653 task representations from in-context learning. In Wanxiang Che, Joyce Nabende, Ekaterina
 654 Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Compu-*
 655 *tational Linguistics: ACL 2025*, pages 6634–6663, Vienna, Austria, July 2025. Association for
 656 Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.345.
 657 <https://aclanthology.org/2025.findings-acl.345/>.

658 Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettle-
 659 moyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proce-*
 660 *edings of the 2024 Conference of the North American Chapter of the Association for Compu-*
 661 *tational Linguistics*, 2024.

662 Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of thoughtlessness? an
 663 analysis of cot in planning. *Advances in Neural Information Processing Systems*, 37:29106–
 664 29141, 2024.

665

666 Sivaramakrishnan Swaminathan, Antoine Dedieu, Rajkumar Vasudeva Raju, Murray Shanahan,
 667 Miguel Lazaro-Gredilla, and Dileep George. Schema-learning and rebinding as mechanisms of
 668 in-context learning and emergence. *Advances in Neural Information Processing Systems*, 36:
 669 28785–28804, 2023.

670 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A ques-
 671 tion answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran,
 672 and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chap-*
 673 *ter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*
 674 *(Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for
 675 Computational Linguistics. doi: 10.18653/v1/N19-1421. <https://aclanthology.org/N19-1421/>.

676

677 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
 678 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
 679 models. In *The Eleventh International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=1PL1NIMMrw>.

680

682 Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few exam-
 683 ples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), June 2020. ISSN 0360-0300.
 684 doi: 10.1145/3386252. <https://doi.org/10.1145/3386252>.

685

686 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
 687 Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-
 688 task language understanding benchmark. *Advances in Neural Information Processing Systems*,
 689 37:95266–95290, 2024.

690 Yuqing Wang and Yun Zhao. Metacognitive prompting improves understanding in large language
 691 models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024*
 692 *Conference of the North American Chapter of the Association for Computational Linguistics: Hu-*
 693 *man Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico,
 694 June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-long.106.
 695 <https://aclanthology.org/2024.nacl-long.106/>.

696 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
 697 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
 698 *neural information processing systems*, 35:24824–24837, 2022.

699

700 Sen Yang, Yunchen Zhang, Leyang Cui, and Yue Zhang. Do prompts solve nlp tasks using natural
 701 language? *CoRR*, abs/2203.00902, 2022. <https://doi.org/10.48550/arXiv.2203.00902>.

702 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
703 React: Synergizing reasoning and acting in language models. In *International Conference on*
704 *Learning Representations*, 2023.

705 Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen,
706 Ligong Han, Kai Xu, Huan Zhang, Dimitris N. Metaxas, and Hao Wang. Token-level uncertainty
707 estimation for large language model reasoning. *ArXiv*, abs/2505.11737, 2025. <https://api.semanticscholar.org/CorpusID:278740334>.

708 710 Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in
711 large language models. In *The Eleventh International Conference on Learning Representations*,
712 2023. <https://openreview.net/forum?id=5NTt8GFjUHkr>.

713 714 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuur-
715 mans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex
716 reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

717 718 Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and
719 Jianbing Shen. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*,
2023.

720 721 Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning
722 Ding, and Bowen Zhou. MedxpertQA: Benchmarking expert-level medical reasoning and un-
723 derstanding. In *Forty-second International Conference on Machine Learning*, 2025. <https://openreview.net/forum?id=IyVcxU0RKI>.

724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

756 **A MATHEMATICAL FORMALIZATION OF SCHEMA-ACTIVATED IN-CONTEXT**
 757 **LEARNING**
 758

759 **A.1 MEMORY: SCHEMAS AND EPISODIC TRACES**
 760

761 We model memory as a bipartite structure linking abstract schemas to multiple episodic examples:

$$762 \quad \mathcal{M} = \{(\mathcal{S}_i, e_j, w_{ij}(t)) \mid \mathcal{S}_i \in \mathcal{S}, e_j \in \mathcal{E}\},$$

763 where

764

- 765 • $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_N\}$ is the set of schemas,
- 766 • $\mathcal{E} = \{e_1, \dots, e_M\}$ is the set of episodic traces (examples),
- 767 • $w_{ij}(t) \in [0, 1]$ is the association strength between schema \mathcal{S}_i and example e_j at time t .

768 Association weights decay over time, modeling episodic forgetting:

$$769 \quad w_{ij}(t) = w_{ij}(0) \cdot \exp(-\lambda t), \quad \lambda > 0.$$

770 The exponential function is an estimate for the forgetting curve, as we want the examples learned
 771 earlier to have less impact.

772 **A.2 PROBLEM REPRESENTATION (SCHEMA FORMATION)**
 773

774 Given an input problem x , the LLM constructs a mental representation (schema):

$$775 \quad \mathcal{S}_x = \mathcal{R}(x)$$

776 where \mathcal{R} is an embedding or representation function.

777 **A.3 PRIOR SCHEMA RETRIEVAL**
 778

779 The model retrieves a schema $\hat{\mathcal{S}}$ from \mathcal{S} :

$$780 \quad \hat{\mathcal{S}} = \arg \max_{\mathcal{S}_i \in \mathcal{S}} \text{sim}(\mathcal{S}_x, \mathcal{S}_i),$$

781 where $\text{sim}(\cdot, \cdot)$ denotes the similarity function, which may be cosine similarity, re-ranking, or another metric.

782 **A.4 EPISODIC EXAMPLES RETRIEVAL**
 783

784 For notational convenience, let $\hat{i} \in \{1, \dots, N\}$ denote the index of $\hat{\mathcal{S}}$ such that $\hat{\mathcal{S}} = \mathcal{S}_{\hat{i}}$. More
 785 formally,

$$786 \quad \hat{i} = \arg \max_{i \in \{1, \dots, N\}} \text{sim}(\mathcal{S}_x, \mathcal{S}_i).$$

787 Afterwards, given the retrieved schema $\hat{\mathcal{S}} = \mathcal{S}_{\hat{i}}$, SA-ICL selects all episodic examples whose (de-
 788 cayed) association to $\hat{\mathcal{S}}$ exceeds a threshold $\tau \in [0, 1]$:

$$789 \quad \hat{\mathcal{E}}_{\tau}(t \mid \hat{i}) = \{e_j \in \mathcal{E} : w_{\hat{i}j}(t) \geq \tau\}.$$

800 **A.5 SCHEMA ACTIVATION (ASSIMILATION / ACCOMMODATION)**
 801

802 The retrieved schema and its selected episodic set guide activation of a new schema for the current
 803 problem:

$$804 \quad \mathcal{S}_{\text{new}} = f(\mathcal{S}_x, \hat{\mathcal{S}}, \hat{\mathcal{E}}_{\tau}(t \mid \hat{i})),$$

805 where f denotes the integration mechanism.

806 Then, *assimilation* and *accommodation* can be conceptualized as follows:

807

- 808 • **Assimilation:** $\mathcal{S}_{\text{new}} \approx \mathcal{S}_x$ when $\hat{\mathcal{S}}$ fits well.
- 809 • **Accommodation:** \mathcal{S}_{new} requires restructuring when fit is poor.

810 A.6 SCHEMA-GUIDED PROBLEM SOLVING
811812 Finally, the LLM produces an output conditioned on the activated schema:
813

814
$$y = \text{LLM}(x, \mathcal{S}_{\text{new}}).$$

815

816 A.7 END-TO-END EQUATION
817818 Combining all steps (schema-first, then thresholded episodic selection), we obtain the following
819 equation:
820

821
$$y = \text{LLM}\left(x, f\left(\mathcal{R}(x), \underbrace{\arg \max_{\mathcal{S}_i \in \mathcal{S}} \text{sim}(\mathcal{R}(x), \mathcal{S}_i)}_{\hat{\mathcal{S}}=\mathcal{S}_i}, \underbrace{\{e_j \in \mathcal{E} : w_{ij}(t) \geq \tau\}}_{\hat{\mathcal{E}}_{\tau}(t|i)}\right)\right).$$

822
823

824 B EXPERIMENTAL SETUP
825826 Table B.1: Experimental setup across model families. "N/A" indicates the parameter count has not
827 been disclosed. Note that embedding and reranker computations were performed on the NVIDIA
828 A40 GPU, with results cached locally.
829

830 Model Family	831 Parameter Count	832 Execution Environment
832 QWen-3	8B	NVIDIA A40 GPU (24GB RAM)
833 LLaMA-3.1	8B	NVIDIA A40 GPU (24GB RAM)
834 Ministeral	8B	API endpoint
835 Gemini 1.5 Flash	N/A	API endpoint
836 GPT-4o Mini	N/A	API endpoint
837 GPT-4o	N/A	API endpoint
838 GPT-5	N/A	API endpoint (subset of experiments)

839
840 C SYNTHETIC DATASET GENERATION
841842 Below are the prompts that we used to generate the synthetic data that are used as our knowledge
843 base for knowledge and schema retrieval mechanisms. The synthetic data are constructed using
844 GPT-4o via the OpenAI API to ensure the GPQA dataset is not included in the LLM's training data.
845 Specifically, we provide exact prompts for each of the three *synthetic similarity* levels: *Essentially*
846 *Same*, *Similar*, and *Different*. Note that we intentionally keep all prompts identical to our experiment
847 setup including format, punctuation, and **typos** to ensure reproducibility.
848849 C.1 *Essentially Same*
850

```
851
852 Consider answer({answer}) and the explanation of solving it({explanation
853   }). this question: {question}, along with its
854
855 Please generate a new question that is distinct from the previous
856   question.
857
858 You should follow the following criteria:
859   - New question requires more knowledge than the provided explanation to
860     be used to answer it.
861   - New question should differ from the given question with a lot of
862     distinctiveness.
863   - Generate a set of new options with only one of them being the correct
864     option to the new question.
865   - Provide three incorrect options, which should be similar to the correct
866     answer
```

864 - Provide a short explanation on how to solve the new question, and the
 865 additional knowledge required to answer the new question.
 866 - **Difficulty:**
 867 The new question should be the similar difficulty to the previous
 868 question.
 869 If a student has the knowledge to answer the previous question, they
 870 should have partial knowledge to answer the new question.
 871 However, the new question should require additional knowledge than
 872 the given question's scope to be answered.
 873 - **Distinctiveness:**
 874 The new question should be distinctive enough to the previous
 875 question, that the student require additional knowledge to solve the
 876 problem.
 877 New question should be unique in its context, and is related to the
 878 previous question in a minimal level.
 879 - **Output Format:**
 880 {question_format}
 881 """

881 C.2 Similar

882
 883 """
 884 Consider this question: {question}, along with its answer({answer}) and
 885 the explanation of solving it({explanation}).

886 Please give me a slightly different question from this example that test
 887 the student's ability to transform their knowledge.
 888

889 You should follow the following criteria:
 890 - The new question only requires the knowledge provided in the
 891 explanation to be used to answer it.
 892 - New question should still differ with a lot of distinctiveness to test
 893 student's use of the same knowledge.
 894 - Generate a set of new options with only one of them being the correct
 895 option to the new question
 896 - Provide three incorrect options, which should be similar to the correct
 897 answer
 898 - Provide a short explanation on how to solve the new question
 899 - **Difficulty:**
 900 The new question should be the similar difficulty to the previous
 901 question.
 902 If a student has the knowledge to answer the previous question, they
 903 should have enough knowledge to answer the new question.
 904 - **Distinctiveness:**
 905 The new question should be distinctive enough to the previous
 906 question, that the student cannot use the same answer.
 907 New question should be unique in its context, but still related to
 908 the previous question.
 909 - **Output Format:**
 910 {question_format}
 911 """

910 C.3 Different

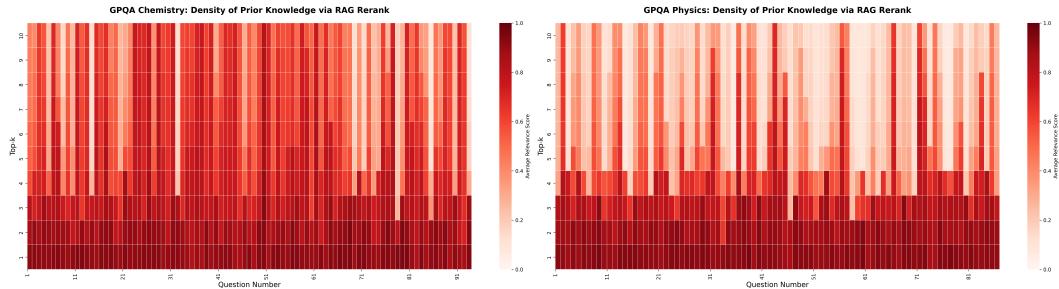
911
 912 """
 913 Consider this question: {question}, along with its answer({answer}) and
 914 the explanation of solving it({explanation}).

915 Please generate a new question that is distinct from the previous
 916 question.
 917

918 You should follow the following criteria:

918 - New question requires more knowledge than the provided explanation to
 919 be used to answer it.
 920 - New question should differ from the given question with a lot of
 921 distinctiveness.
 922 - Generate a set of new options with only one of them being the correct
 923 option to the new question.
 924 - Provide three incorrect options, which should be similar to the correct
 925 answer
 926 - Provide a short explanation on how to solve the new question, and the
 927 additional knowledge required to answer the new question.
 928 - **Difficulty:**
 929 The new question should be the similar difficulty to the previous
 930 question.
 931 If a student has the knowledge to answer the previous question, they
 932 should have partial knowledge to answer the new question.
 933 However, the new question should require additional knowledge than
 934 the given question's scope to be answered.
 935 - **Distinctiveness:**
 936 The new question should be distinctive enough to the previous
 937 question, that the student require additional knowledge to solve the
 938 problem.
 939 New question should be unique in its context, and is related to the
 940 previous question in a minimal level.
 941 - **Output Format:**
 942 {question_format}
 943
 944

942 D HEATMAPS FOR KNOWLEDGE DENSITY IN LATENT SIMILARITY



945 Figure D.1: The heatmaps for knowledge density for GPQA-Chemistry and GPQA-Physics
 946
 947

948 Knowledge density is defined as the similarity between prior knowledge and a target question.
 949 Similarity is quantified by converting prior knowledge into mental representations and employing
 950 cross-encoder rerankers to compute relevance scores. The knowledge density heatmaps Figure D.1
 951 visualize marginal relevance scores across top-k retrieved examples for each question, revealing
 952 domain-specific patterns in knowledge sparsity. Our analysis demonstrates that chemistry maintains
 953 consistently high relevance scores across retrieved examples, while physics exhibits significantly
 954 greater sparsity — with relevance scores decreasing 68% compared to chemistry's 40% decrease.
 955 These findings underscore that effective schema-based knowledge transfer depends on inherent con-
 956 ceptual coherence within domains rather than universal retrieval mechanisms.
 957

958 D.1 LATENT KNOWLEDGE DENSITY METHODOLOGY

959 We define the latent knowledge density function $\rho : Q \times K \rightarrow [0, 1]$, where Q represents the set
 960 of questions and $K = \{1, 2, \dots, k_{max}\}$ denotes retrieval ranks. In our case, $k_{max} := 10$. For each
 961 question $q \in Q$ and rank $k \in K$, $\rho(q, k)$ measures the relevance score between q and its k -th
 962 most similar prior example, computed using Cohere's Rerank 3.5 cross-encoder architecture. The
 963 aggregate density metric $\bar{\rho}(q) = \frac{1}{k_{max}} \sum_{k=1}^{k_{max}} \rho(q, k)$ quantifies overall knowledge availability for
 964

972 question q , where higher values indicate robust transfer potential and lower values reveal knowl-
 973 edge sparsity. We define $\sigma \in [0, 1]$ as the relevance threshold hyperparameter to quantify what we
 974 consider quality knowledge. For this investigation, we arbitrarily set $\sigma := 0.5$.
 975

976 D.2 HEATMAP VISUALIZATION AND INTERPRETATION 977

978 Figure D.1 visualizes $\rho(q, k)$ for GPQA-Chemistry and GPQA-Physics datasets as heatmaps, where
 979 each cell (i, j) represents $\rho(q_i, k_j)$ with q_i denoting the i -th question (horizontal axis) and $k_j \in$
 980 $\{1, 2, \dots, k_{max}\}$ the retrieval rank (vertical axis). Color intensity maps linearly to ρ values, from
 981 light shades ($\rho \rightarrow 0$) to dark red ($\rho \rightarrow 1$), symbolizing low to high relevance scores respectively.

982 Chemistry questions maintain $\rho_{chem}(q, k) > \sigma$ for most $(q, k) \in Q \times K$ pairs even at $k =$
 983 k_{max} , indicating persistent semantic similarity with small $\Delta\rho = \rho(q, k) - \rho(q, k + 1)$ for all
 984 $k \in \{1, 2, \dots, k_{max} - 1\}$. Physics exhibits rapid density decay with $\rho_{phys}(q, k) < 0.3$ for many
 985 questions by $k = 5$, revealing sparse knowledge availability beyond initial retrievals.
 986

987 D.3 DOMAIN SPARSITY AND KNOWLEDGE SPACE ISOLATION 988

989 Defining decay rate as $\theta(k) = \frac{\rho(q, k)}{\rho(q, 1)}$, we observe $\bar{\theta}_{chem}(k_{max}) = 0.60$ for chemistry versus
 990 $\bar{\theta}_{phys}(k_{max}) = 0.32$ for physics, where the bar notation indicates domain averaging. This repre-
 991 sents $\bar{\rho}_{chem}(q, k_{max}) \approx 0.6 \cdot \bar{\rho}_{chem}(q, 1)$ netting a 40% decrease, compared to $\bar{\rho}_{phys}(q, k_{max}) \approx$
 992 $0.32 \cdot \bar{\rho}_{phys}(q, 1)$ which yields a 68% decrease. Chemistry maintains $\bar{\rho}_{chem}(q, k) > \sigma$ throughout all
 993 $k \in \{1, 2, \dots, k_{max}\}$, while physics shows a domain-specific density gradient $\nabla_k \rho$ approximately
 994 twice as steep.
 995

996 We hypothesize that this disparity stems from the inherent conceptual fragmentation of physics
 997 across fields like quantum mechanics and general relativity — each possessing specialized theoreti-
 998 cal vocabularies that create isolated knowledge regions within the latent space. The intercluster den-
 999 sity satisfies $\rho(q_i, e_j) \ll \sigma$ when q_i and the episodic trace e_j belong to different physics subdomains.
 1000 For example, specialized concepts inherent to quantum mechanics, such as quantum entanglement
 1001 and wave functions, yield $\rho \approx 0$ when matched against general relativity examples, forming disjoint
 1002 clusters and thus leading to poorer schema activations.
 1003

1004 These findings demonstrate that effective schema-based knowledge transfer depends critically on
 1005 domain density structure $\rho(q, k)$. High-density regimes where $\bar{\rho} > \sigma$ support extensive retrieval
 1006 sets, while sparse domains with rapidly decaying $\rho(q, k)$ require adaptive mechanisms for $\rho \rightarrow 0$
 1007 when $k > k_{threshold}$.
 1008

1009 E DOMAIN-AGNOSTIC SCHEMA TEMPLATE 1010

1011 When generating mental representations or schemas, LLMs were guided for each attribute of the
 1012 schema.
 1013

1014 **"""Drawing on schema theory from cognitive psychology, think about a high
 1015 -level abstraction (schema) of the problem to guide your reasoning.
 1016 Your ultimate goal is to select the most appropriate answer.:**

1017 **Below is the template for the schema you need to fill out:**

1018 **Broad Category:**

1019 **Identify the overarching subject and general category to which the
 1020 problem belongs.**

1021 **Refinement:**

1022 **Describe further details or specific aspects that narrow down the broad
 1023 category.**

1024 **Specific Scope:**

1025 **Define the precise focus or context of the problem within the refined
 1026 category.**

```

1026 Goal:
1027 Clearly state the objective or intended outcome of solving the problem.
1028
1029 Finally, summarize the schema in a few sentences to help students grasp
1030 the key points. The problem you need to abstract is as follows:""""
1031
1032 The LLM will then generate a JSON-like object that represents a concrete schema for a specific
1033 question:
1034
1035 {
1036     "schema": {
1037         "broad_category": str,
1038         "refinement": str,
1039         "specific_scope": str,
1040         "goal": str
1041     },
1042     "summary": str
1043 }
1044
1045
1046
1047 F A FULL CONVERSATION HISTORY DURING THE INFERENCE TIME WITH
1048 GPT-4O MINI
1049
1050 Note that all three methods use the same dynamic response JSON format:
1051
1052 {
1053     "name": "DynamicResponse",
1054     "strict": true,
1055     "schema": {
1056         "$defs": {
1057             "AnswerEnum": {
1058                 "enum": [
1059                     "14",
1060                     "12",
1061                     "10",
1062                     "11"
1063                 ],
1064                 "title": "AnswerEnum",
1065                 "type": "string"
1066             }
1067         },
1068         "properties": {
1069             "reasoning": {
1070                 "title": "Reasoning",
1071                 "type": "string"
1072             },
1073             "final_answer": {
1074                 "$ref": "#/$defs/AnswerEnum"
1075             }
1076         },
1077         "required": [
1078             "reasoning",
1079             "final_answer"
1080         ],
1081         "title": "DynamicResponse",
1082         "type": "object",
1083         "additionalProperties": false
1084     }
1085 }
1086
1087
1088 F.1 SCHEMA-ACTIVATED IN-CONTEXT LEARNING
1089
1090 System Message:

```

```

1080     """Drawing on schema theory from cognitive psychology, think about a high
1081     -level abstraction (schema) of the problem to guide your reasoning.
1082     Your ultimate goal is to select the most appropriate answer.:
1083     Below is the template for the schema you need to fill out:
1084     Broad Category:
1085     Identify the overarching subject and general category to which the
1086     problem belongs.
1087     Refinement:
1088     Describe further details or specific aspects that narrow down the broad
1089     category.
1090     Specific Scope:
1091     Define the precise focus or context of the problem within the refined
1092     category.
1093     Goal:
1094     Clearly state the objective or intended outcome of solving the problem.
1095     Finally, summarize the schema in a few sentences to help students grasp
1096     the key points. The problem you need to abstract is as follows:"""
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

```

User Messages:

```

[ {
  "role": "user",
  "content": "Question: trans-cinnamaldehyde was treated with
  methylmagnesium bromide, forming product 1.\n\n1 was treated with
  pyridinium chlorochromate, forming product 2.\n\n3 was treated with (dimethyl(oxo)-16-sulfaneylidene)methane in DMSO at elevated
  temperature, forming product 3.\n\nhow many carbon atoms are there in
  product 3?\nCandidates: \n1. 14\n2. 12\n3. 10\n4. 11"
},
{
  "role": "assistant",
  "content": "#### Schema:\n#### broad_category:\nOrganic Chemistry \u2192 Reaction Mechanisms \u2192 Organic Synthesis\n\n#### refinement:\nThis problem involves a series of organic reactions starting from trans-cinnamaldehyde and leading to a final product through specific reagents and conditions.\n\n#### specific_scope:\nThe focus is on determining the number of carbon atoms in the final product (product 3) after a sequence of reactions involving trans-cinnamaldehyde, methylmagnesium bromide, pyridinium chlorochromate, and a sulfaneylidene compound.\n\n#### goal:\nCalculate the total number of carbon atoms in product 3 after completing the series of reactions.\n\n#### Summary:\nThis problem involves a series of organic reactions that transform trans-cinnamaldehyde into a final product through various reagents and conditions. The task is to track the carbon atoms throughout the reactions to determine the total number in the final product (product 3). By analyzing the structure of each intermediate and the final product, we can ascertain the number of carbon atoms present. The goal is to arrive at the correct count of carbon atoms in product 3."
},
{
  "role": "user",
  "content": "Now that you've seen a similar question along with its schema and solution, reflect on how you could refine or improve your previously generated schema for the **current** question. Focus on alignment in categories, scope specificity, and consistency of abstraction.\n## Question:\nQuestion: Trans-cinnamaldehyde undergoes a reaction with methylmagnesium bromide to yield compound 1. Compound 1 is then oxidized using pyridinium chlorochromate to form compound 2. Subsequently, compound 2 reacts with (dimethyl(oxo)-\u03bb6-sulfaneylidene)methane in DMSO at high temperature to produce compound 3. How many carbon atoms are present in compound 3?\nCandidates: \n1. 12\n2. 10\n3. 9\n4. 11\n\n## Answer to this question is: 11\n\n## Below is the schema and summary of this"
}

```

```

1134 question: \n#### Schema:\n#### broad_category:\nChemistry \u2192
1135 Organic Chemistry \u2192 Reaction Mechanisms\n\n#### refinement:\n
1136 This problem involves a series of organic reactions including
1137 Grignard reactions, oxidation, and a reaction with a sulfonyl
1138 compound.\n\n#### specific_scope:\nThe focus is on determining the
1139 number of carbon atoms in the final product (compound 3) after a
1140 sequence of reactions starting from trans-cinnamaldehyde.\n\n#### goal:\nCalculate the total number of carbon atoms in compound 3 after
1141 all specified reactions have been completed.\n\n#### Summary:\nThis
1142 problem involves a series of organic reactions that transform trans-
1143 cinnamaldehyde through various steps, including a Grignard reaction,
1144 oxidation, and a reaction with a sulfonyl compound. The objective is
1145 to determine the number of carbon atoms in the final product,
1146 compound 3. By analyzing the reactions and counting the carbon atoms
1147 at each stage, we find that compound 3 contains 11 carbon atoms."
1148 },
1149 {
1150     "role": "assistant",
1151     "content": "#### Schema:\n#### broad_category:\nOrganic Chemistry \u2192 Synthetic Transformations \u2192 Carbon Skeleton Accounting\n\n#### refinement:\nMulti-step sequence featuring: (1) Grignard addition to an \u03b1,\u03b2-unsaturated aldehyde (trans-cinnamaldehyde), (2) PCC oxidation of the resulting alcohol to a ketone, and (3) Corey\u2013Chaykovsky-type methylene transfer using dimethylsulfoxonium methylide in DMSO at elevated temperature. Emphasis on how each step affects carbon count.\n\n#### specific_scope:\nTrack carbon atoms starting from trans-cinnamaldehyde (C9). Step 1: MeMgBr adds a methyl to the carbonyl carbon (allylic secondary alcohol) \u2192 +1 carbon. Step 2: PCC oxidizes alcohol to ketone \u2192 0 carbon change. Step 3: Dimethyl(oxo)-\u03bb6-sulfanylidene)methane (dimethylsulfoxonium methylide) transfers a methylene to the carbonyl (Corey\u2013Chaykovsky epoxidation/methylenation) \u2192 +1 carbon. Determine the total carbons in product 3.\n\n#### goal:\nUse stepwise carbon accounting across the reaction sequence to identify the number of carbon atoms in product 3 and select the correct option.\n\n#### Summary:\nThis is a carbon-tracking problem across a three-step synthesis from trans-cinnamaldehyde. Only the Grignard addition and the sulfur ylide methylene transfer change the carbon count, each adding one carbon; the PCC oxidation does not. Applying this schema leads to the correct carbon total in the final product."
1152 },
1153 {
1154     "role": "user",
1155     "content": "Now that you've refined your schema for the current
1156 question, select the most appropriate answer. "
1157 }
1158 ]

```

1176 F.2 ONE-SHOT

1177 System Message:

```

1178 """Select the most appropriate answer. You will be shown a sample
1179 question with its correct answer.
1180 Then, answer a new question of the same type."""
1181

```

1182 User Messages:

```

1183 [
1184     {
1185         "role": "user",
1186         "content": "Example Question:\nQuestion: Trans-cinnamaldehyde
1187 undergoes a reaction with methylmagnesium bromide to yield compound

```

```

1188
1189 1. Compound 1 is then oxidized using pyridinium chlorochromate to
1190 form compound 2. Subsequently, compound 2 reacts with (dimethyl(oxo)
1191 -\u03bb6-sulfaneylidene)methane in DMSO at high temperature to
1192 produce compound 3. How many carbon atoms are present in compound 3?\n
1193 Candidates: \n1. 12\n2. 10\n3. 9\n4. 11\nAnswer: 11"
1194 },
1195 {
1196 "role": "user",
1197 "content": "Now try a similar question:\nQuestion: trans-
1198 cinnamaldehyde was treated with methylmagnesium bromide, forming
1199 product 1.\n\nn1 was treated with pyridinium chlorochromate, forming
1200 product 2.\n\nn3 was treated with (dimethyl(oxo)-16-sulfaneylidene)
1201 methane in DMSO at elevated temperature, forming product 3.\n\nnhow
1202 many carbon atoms are there in product 3?\nCandidates: \n1. 14\n2.
1203 12\n3. 10\n4. 11"
1204 }
1205
1206
1207
1208 F.3 ONE-SHOT + CoT
1209
1210 System Message:
1211
1212
1213 """Select the most appropriate answer. You will be shown a sample
1214 question with its correct answer.
1215 Then, answer a new question of the same type."""
1216
1217
1218 User Messages:
1219
1220
1221
1222 [
1223 {
1224 "role": "user",
1225 "content": "Example Question:\nQuestion: Trans-cinnamaldehyde
1226 undergoes a reaction with methylmagnesium bromide to yield compound
1227 1. Compound 1 is then oxidized using pyridinium chlorochromate to
1228 form compound 2. Subsequently, compound 2 reacts with (dimethyl(oxo)
1229 -\u03bb6-sulfaneylidene)methane in DMSO at high temperature to
1230 produce compound 3. How many carbon atoms are present in compound 3?\n
1231 Candidates: \n1. 12\n2. 10\n3. 9\n4. 11\nAnswer: 11"
1232 },
1233 {
1234 "role": "user",
1235 "content": "Now try a similar question:\nQuestion: trans-
1236 cinnamaldehyde was treated with methylmagnesium bromide, forming
1237 product 1.\n\nn1 was treated with pyridinium chlorochromate, forming
1238 product 2.\n\nn3 was treated with (dimethyl(oxo)-16-sulfaneylidene)
1239 methane in DMSO at elevated temperature, forming product 3.\n\nnhow
1240 many carbon atoms are there in product 3?\nCandidates: \n1. 14\n2.
1241 12\n3. 10\n4. 11\nPlease think step by step."
1242 }
1243 ]
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
```

1242 **G GPT-5’s PERFORMANCE ON HUMANITY’s LAST EXAM**
1243

1262 Figure G.2: Performance of GPT-5 on the chemistry, physics, biology multi-choice questions from
1263 the Humanity’s Last Exam (HLE) benchmark dataset (Phan et al., 2025) (N=191, we filtered out
1264 questions with images), when provided with the *Essentially Same* examples. We generated the
1265 synthetic data as well using the same procedures described in Section C.

1266
1267 When GPT-5 was introduced, we used the model to run the same experiments. However, we re-
1268 alized that when provided with *Essentially Same* examples, GPT-5 reached near 100% accuracy
1269 in the GPQA dataset, eliminating the need for explicit reasoning modules introduced by SA-ICL.
1270 Therefore, we used a more challenging dataset, Humanity’s Last Exam (HLE), to repeat the same
1271 experiments using *Essentially Same* examples, and saw that SA-ICL results in a 7% improvement in
1272 accuracy compared to standard One-Shot. This gain is particularly notable given the already strong
1273 baseline accuracy of 84%.

1274 Note that this subset of HLE includes 136 biology and medicine questions, 34 physics questions, and
1275 21 chemistry questions. We acknowledge that most of the questions are in the biology and medicine
1276 sub-categories, which is a domain that was not tested in the main experiments shown in Section 5.3.
1277

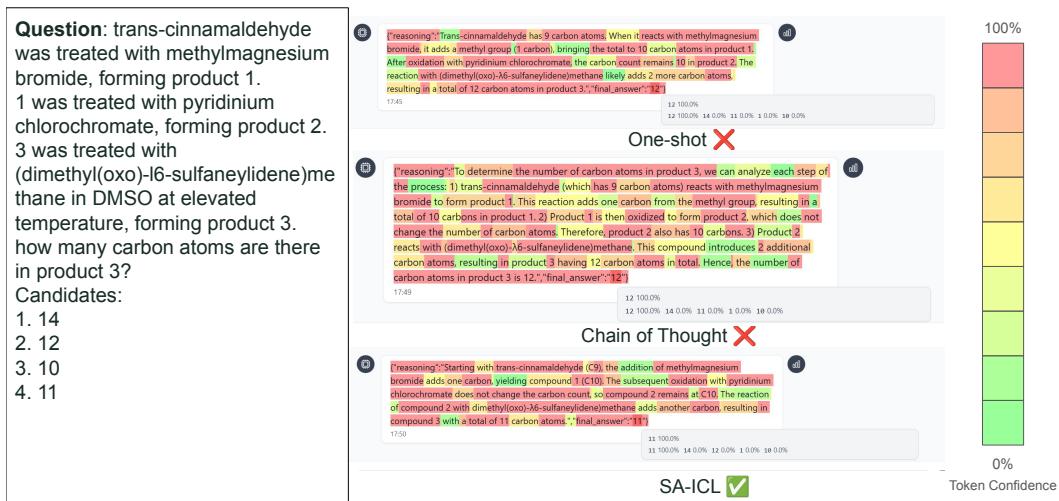
1296 H INTERPRETABILITY EXPERIMENTS

1298 H.1 REPORT ON TOKEN COUNTS AND CORRECTNESS FOR THE FIRST 10 GPQA-Chemistry 1299 QUESTIONS

1301 The interpretability experiments are conducted via logit-centric methods to get the overall confi-
1302 dence of LLM generation by retrieving the token-level log likelihood (Zhang et al., 2025) from the
1303 OpenAI chat completions API for each of our tested ICL prompting strategies. We implemented
1304 a customized chat interface to display the log-probability in colored texts using the color schema
1305 provided in the legend in Figure H.3 for better visualization and interaction.

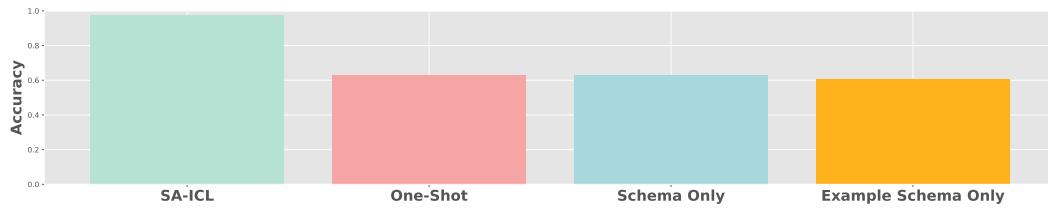
1306 From Table 3, we observed that SA-ICL achieves the highest accuracy in these questions while
1307 simultaneously generating fewer reasoning tokens compared to One-Shot + CoT, except for the
1308 sixth question.

1310 H.2 VISUALIZATION OF THE LOG-PROBABILITIES OF EACH TOKEN IN THE SOLVER'S 1311 OUTPUT



1329 Figure H.3: The probabilities of each token in the problem solver output. Note that only SA-ICL
1330 confidently gave the correct answer for this question, while pure One-Shot or One-Shot + CoT
1331 confidently gave the incorrect answer.

1350 **I ABLATION STUDY ON THE SCHEMA ACTIVATION**
 1351
 1352



1360 Figure I.4: GPT-4o Mini solving the *GPQA-Physics* questions with examples that are *Essentially*
 1361 *Same*.

1362
 1363 To further investigate in isolation the effectiveness of schema activation within our framework,
 1364 we conducted a controlled ablation study comparing four distinct conditions: SA-ICL, One-Shot,
 1365 Schema Only, and Example Schema only.

1366
 1367 **I.1 DESIGN AND NATURE OF ABLATION STUDY**
 1368

1369 This ablation analysis was designed to determine whether the performance improvements observed
 1370 in SA-ICL stem from the schema activation mechanism itself or merely from the presence of ab-
 1371 stracted examples. In particular, we are interested in the case when prior examples are of high
 1372 quality, and therefore conducted this study by using GPT-4o Mini to solve questions in the *GPQA-*
 1373 *Physics* subset with prior examples that are *Essentially Same* in terms of synthetic similarity. The
 1374 Schema Only condition provided models with abstract problem schemas without corresponding
 1375 episodic examples, while Example Schema Only presented abstracted schemas derived from exam-
 1376 ples but without the schema activation process that is integral to the full SA-ICL procedure.

1377
 1378 **I.2 INTERPRETATION OF ABLATION RESULTS**

1379 Figure I.4 presents the comparative accuracy results across all four scenarios. The complete SA-
 1380 ICL framework demonstrated clear superiority over the other three approaches, with a near 40%
 1381 accuracy boost over the other three conditions. In fact, neither Schema Only nor Example Schema
 1382 Only result in a clear benefit over One-Shot, which acts as a baseline for this ablation study. This
 1383 observation provides direct evidence that the schema activation mechanism constitutes the critical
 1384 component driving the effectiveness of our in-context learning framework.

1385
 1386 **I.3 SCHEMA DORMANCY**
 1387

1388 The results of our ablation study strongly suggest that the mere presence of an abstracted schema
 1389 within the prompt context is insufficient to improve model performance. The sheer significance of
 1390 schema activation leads to a phenomenon which we coin *schema dormancy*. This term describes a
 1391 state where a refined schema tailored to a specific task only exists as passive surface-level contextual
 1392 information rather than an active cognitive framework that guides the language model’s reasoning
 1393 process. When the explicit activation step of the SA-ICL framework is omitted — as observed in
 1394 the Schema Only and Example Schema Only conditions — the LLM fails to effectively integrate
 1395 the abstracted reasoning structure into its problem-solving approach. This finding highlights that
 1396 language models do not implicitly adopt and utilize abstract schemas without an explicit technique
 1397 to activate this otherwise dormant state.

1398 **THE USE OF LARGE LANGUAGE MODELS**
 1399

1400 The authors used large language models such as ChatGPT to assist with paper writing, mostly for
 1401 polishing the text, and after which the authors did a thorough check to ensure the polished paper
 1402 faithfully delivered the authors’ messages.