EXPLORA: Efficient Exemplar Subset Selection for Complex Reasoning

Kiran Purohit
IIT Kharagpur
kiran.purohit@kgpian.iitkgp.ac.in

Venktesh V
TU Delft
v.viswanathan-1@tudelft.nl

Raghuram Devalla
IIT Kharagpur
devallaraghuram@gmail.com

Krishna Mohan Yerragorla
IIT Kharagpur
krishnamohanyerragorla@gmail.com

Sourangshu Bhattacharya
IIT Kharagpur
sourangshu@cse.iitkgp.ac.in

Avishek Anand
TU Delft
avishek.anand@tudelft.nl

Abstract

Answering reasoning-based complex questions over text and hybrid sources, including tables, is a challenging task. Recent advances in large language models (LLMs) have enabled in-context learning (ICL), allowing LLMs to acquire proficiency in a specific task using only a few demonstration samples (*exemplars*). A critical challenge in ICL is the selection of optimal exemplars, which can be either task-specific (static) or test-example-specific (dynamic). Static exemplars provide faster inference times and increased robustness across a distribution of test examples. In this paper, we propose an algorithm for static exemplar subset selection for complex reasoning tasks. We introduce EXPLORA, a novel exploration method designed to estimate the parameters of the scoring function, which evaluates exemplar subsets without incorporating confidence information. EXPLORA significantly reduces the number of LLM calls to ~11% of those required by state-of-the-art methods and achieves a substantial performance improvement of 12.24%. We open-source our **code and data**[1].

1 Introduction

Answering complex questions that require multi-step reasoning (Chen et al., 2022b; Lu et al., 2023a; Ling et al., 2017; Roy and Roth, 2016; Roy and Anand, 2022; Venktesh et al., 2024) over structured and unstructured sources is an active research area with applications in finance, law, fact-checking and healthcare (Wang et al., 2023a; Zhang et al., 2021). Unlike fine-tuning task-specific models (Chiang and Chen, 2019; Amini et al., 2019; Ling et al., 2017; Roy and Roth, 2016; Geva et al., 2020; Cobbe et al., 2021), recent advances in Large Language Models (LLMs) have paved the way for new approaches that employ in-context learning (ICL) (Wei et al., 2023) to solve complex reasoning problems. This approach focuses on choosing a small number of demonstration examples to be used as input prompts to LLMs.

An effective way to tackle complex reasoning problems is to use *chain-of-thought* (COT) prompting, which adds *hand-crafted natural language rationales* as stepwise solutions to the prompts, resulting in the triplet `(input, rationale, output)` (Wei et al., 2023). In this work, we refer to this triplet as an **exemplar**. A limitation of the COT-based method for reasoning tasks is the tedious and non-scalable manual effort required in selecting the rationales or *exemplars* (Lu et al., 2022; Zhao et al., 2021; Chang and Jia, 2023). To address these limitations, both static and dynamic approaches for automatic exemplar selection have been proposed (Ye et al., 2023a; Lu et al., 2022; Rubin et al., 2022; Fu et al., 2023). Some approaches require annotated data and training of multiple models for exemplar selection (Lu et al., 2023b; Ye et al., 2023a). Dynamic exemplar selection methods often involve additional computational costs because they select exemplars during query time, necessitating extensive query encoding and dynamic exploration of the search space. In contrast, static exemplar selection pre-selects a small subset of exemplars, which are used during LLM inference. Prior exemplar selection methods do not capture interactions between the exemplars in the selected set (Li and Qiu, 2023). Additionally, current static selection approaches (Li and Qiu, 2023) are characterized by a large number of LLM calls, which are computationally expensive.

In this work, we propose EXPLORA, a novel *static exemplar subset selection* method that selects multiple low-loss exemplar subsets (overview in Figure 1). Our method is designed based on two hypotheses: (1) An effective exemplar selection algorithm for ICL should model the end-to-end ICL process, and (2) Prompt generators (refer Sec-

[1] https://github.com/kiranpurohit/EXPLORA

tion 3.1), which generate prompts using multiple exemplar subsets can be used to enhance the effectiveness of static ICL predictors for complex reasoning tasks. Following these hypotheses, we model the problem of static exemplar selection for *In-context Complex Reasoning* (ICCR) tasks as a novel top-l *exemplar subset-selection problem*. We use a linear model of similarity with validation examples for modeling the loss incurred by exemplar subsets. We propose a novel sampling-based bandit algorithm for simultaneously estimating the parameters of the loss model and identifying the top-l exemplar subsets, while incurring a low number of calls to the LLM (which corresponds to a low sample complexity of the bandit algorithm). Our approach implicitly captures the interactions between the exemplars in the subset by scoring subsets. We conduct extensive experiments across multiple reasoning-based QA tasks. Our results indicate that EXPLORA outperform both static and dynamic exemplar selection baselines by 12.24% and 45.45% respectively (Table 1), while reducing the number of LLM calls to $\sim 11\%$ of the state-of-the-art (Li and Qiu, 2023) (Figure 2).

Contributions. The contributions of our work are:

(1) We propose a novel top-l exemplar-subset selection approach, EXPLORA, for end-to-end in-context learning of complex reasoning tasks by approximating the loss for a given exemplar subset using a scoring function.

(2) We introduce a novel sampling-based bandit algorithm for efficiently learning the parameters of the scoring function and estimating the top-l exemplar subsets.

(3) We demonstrate that the exemplars selected by EXPLORA on smaller LLMs can be well *transferred* to larger LLMs (Table 2), reducing the cost incurred by larger LLMs for exemplar selection.

(4) We show that exemplars selected by EXPLORA are more *robust* compared to baselines in task performance (Table 3).

2 Related Work

While many existing techniques for complex reasoning tasks involve fine-tuning of specialized models (Chiang and Chen, 2019; Amini et al., 2019; Chen et al., 2020), these approaches require access to the model parameters. Recent developments in language models have introduced few-shot prompting approaches (Brown et al., 2020a; Wei et al., 2022) through ICL (Wei et al., 2023; Wang et al., 2023b; Kojima et al., 2023; Chen et al., 2022a) and COT in complex reasoning tasks (Ling et al., 2017; Cobbe et al., 2021; Brown et al., 2020b; Shin et al., 2020; V et al., 2023). However, a major drawback of these approaches is the need for manual selection of exemplars, which is tedious and non-scalable. Moreover, ICL is sensitive to the sample order (Lu et al., 2022), dataset, task, and models (Zhao et al., 2021) (Su et al., 2022), making optimal exemplar selection essential for stable task performance.

Exemplar-Selection for ICL: Several automated exemplar selection methods have been proposed to eliminate the need for manual selection. These include reinforcement learning-based approaches (Zhang et al., 2022; Lu et al., 2023b), Determinantal Point Processes (Ye et al., 2023a), Low Rank approximation (DQ-LoRe) (Xiong et al., 2023) and constrained optimization (Tonglet et al., 2023), which are effective for reasoning tasks. Additionally, alternative learning-free methods for diverse exemplar selection, such as similarity-based (Rubin et al., 2022), complexity-based (Fu et al., 2023) and MMR (Ye et al., 2023b) have also been proposed.

However, existing dynamic exemplar selection methods incur additional computational costs during inference. To address this, a small, representative set of exemplars can be selected for ICL. Unlike coreset selection methods (Guo et al., 2022), which rely on gradient-based model updates for traditional deep learning, ICL performs the target tasks without any parameter updates.

To the best of our knowledge, there has been very little research in this area, with the closest work being LENS (Li and Qiu, 2023). However, LENS relies on LLM's output probabilities and thus cannot be extended to black-box LLMs. In this work, we propose a novel and robust approach for static exemplar selection that is applicable to black-box and also other open language models.

3 EXPLORA: Model-based Exploration for Exemplar Subset Selection

In black-box models, we cannot access the parameters of the LLMs or compute the gradient of the loss with respect to these parameters. Additionally, intermediate representations or generative probability scores from LLMs cannot be utilized for scoring exemplar subsets. To overcome these challenges, we introduce a novel approach to effectively select exemplar subsets without relying on the parameters of the LLMs. Section 3.1 formally describes the

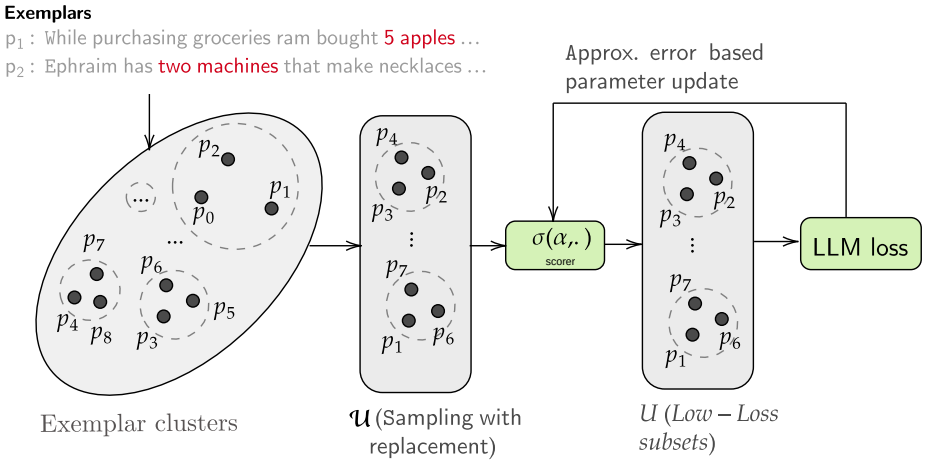

Figure 1: **Overview of EXPLORA:** Initially, set U is randomly selected from set $\mathcal{U}$. In each iteration, parameters of the scoring function $\sigma(\alpha, .)$ are computed by minimizing a loss function. σ guides the selection of the subset from $\mathcal{U} \setminus U$ with the lowest loss, which is then used to update U. This iterative updating process ensures U maintains low-loss subsets, leading to a more accurate estimation of α in subsequent iterations.

components of ICL. Section 3.2 formulates the loss-model for exemplar subset selection. Section 3.3 motivates and describes the EXPLORA algorithm.

3.1 ICL for Complex Reasoning

In-context learning (ICL) leverages LLMs to acquire proficiency in a specific task using only a few exemplars, without updating the model's parameter. During this process, the LLM is provided with a prompt that includes *input-rationale-output* triplet, referred to as *exemplars*, which demonstrate the task to the LLM. Due to the financial and performance costs associated with large contexts, providing all n training examples is impractical. Hence, exemplar selection methods curate a few exemplars that maximize overall accuracy. Formally, let $\mathcal{X} = \{x_i, z_i, y_i\}_{i=1}^{n}$ denote the set of all n training examples (potential exemplars), and let x_{test} be a test input. The goal is to predict the test output y_{test}. Let $S \subseteq \mathcal{X}$ denote a subset of k exemplars used for predicting x_{test}. The prompt P is constructed as: $P = [S, x_{test}] = [(x_{i_1}, z_{i_1}, y_{i_1}), ..., (x_{i_k}, z_{i_1}, y_{i_k}), x_{test}]$, The end-to-end ICL process can be described as a composition of two steps: (1) response generator f, and (2) post-processing (decoding) δ. The response generator f generates multiple responses, r from the probability distribution $\mathbb{P}_{LLM}$. The post-processing step δ is applied to the LLM-generated response $f(P)$ to extract the task-specific output $\hat{y}_{test}$.

$$\hat{y}_{test} = \delta(f([S, x_{test}])); \; f(P) = \mathcal{G}\left(\mathbb{P}_{LLM}(r|P)\right) \quad (1)$$

Commonly used post-processing strategies include regular-expression matching (δ_{regex}) and self-consistency (δ_{SC}) (Wang et al., 2023c).

Reasoning problems (Ho et al., 2020; Ling et al., 2017) are especially challenging due to the complex relationship between the exemplars and the LLM's ability to perform multi-step reasoning tasks. It has been demonstrated that providing rationales elucidating the reasoning steps improves LLM performance compared to state-of-the-art approaches (Wei et al., 2023; Fu et al., 2023). Therefore, selecting exemplars with appropriate rationales tailored to the complex reasoning task is crucial (Xiong et al., 2023; Tonglet et al., 2023). However, generic exemplar selection approaches for ICL do not explicitly model this role of the rationales and interactions between exemplars. End-to-end modeling of the entire ICL process (Eq 1) is essential for capturing the complex role of the rationales. In this work, we propose implicitly modeling the relationship between exemplars and explicitly modeling their relation to the LLM's performance by scoring subsets of exemplars (see section 3.2). This leads to formulation for the *In-context Complex Reasoning* (ICCR) problem as an exemplar-subset selection problem.

A disadvantage of using a single subset of exemplars is that the prompt may not capture all diverse aspects of a given task. This can be addressed using a prompt generator $\pi(S_1, ..., S_l, x_{test})$, which uses multiple subsets of exemplars, $S_1, ..., S_l$ and a test input x_{test} to create a prompt P, improving overall performance and robustness compared to

using a single exemplar subset. For example, one can use a similarity-based prompt generator π_{KNN}, which selects semantically nearest neighbor subset, or a diversity-based prompt generator π_{MMR} (the details are described in section 4). In this revised framework, the entire output generation process can be described as:

$$\hat{y}_{test} = \delta(f(P));\ P = \pi(S_1, ..., S_l, x_{test}) \quad (2)$$

In this setup, we are interested in finding a set $U = \{S_1, ..., S_l\}$ of subsets such that the corresponding prompt P generated by the prompt generator π minimizes the total validation loss. Let $\mathcal{V}$ be the set of m *validation examples* $\{u_i, v_i\}_{i=1}^m$, where u_i and v_i represents the *input* and the *output* of the i^{th} validation example respectively. We define the validation loss for a prompt generator $\pi(U)$ with a set of exemplar subsets U as:

$$L(\pi, U, \mathcal{V}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(v_i \neq \delta(f(\pi(U, u_i)))) \quad (3)$$

Hence, we define the problem of top-l exemplar subset selection for ICCR as:

Definition 3.1 (ICCR). *Let $\mathcal{V} = \{u_i, v_i\}_{i=1}^m$ be a set of validation examples, and $U = \{S_1, ..., S_l\}$ denote a set of subsets S_i of size k. The problem of exemplar subset selection can be described as:*

$$U^* = \underset{U=\{S_1,...,S_l\}}{\arg\min}\ L(\pi, U, \mathcal{V}) \quad (4)$$

, $S_i \subseteq \mathcal{S}$, $\mathcal{S}$ is the set of all subsets S, $|S| = k$.

3.2 Loss model for top-l Exemplar Subset Selection

There are two main challenges in efficiently solving the exemplar subset selection problem (Eq 4): (1) the set of all exemplar subsets $\mathcal{S}$ can be very large leading to prohibitive time-complexity. For instance, $n = 5000$ training examples, and a prompt size of $k = 5$ examples lead to C_{5000}^5 (approximately $2.5 * 10^{16}$) exemplar subsets. (2) a naive calculation of loss (Eq 3) for each exemplar subset S involves m calls ($m \sim 1000$) to the LLM, which can be expensive (both computationally and financially). We address the first issue in section 3.3. We propose to address the second issue by building a *scoring function* (σ) for the subsets $S \in \mathcal{S}$, which can be used to calculate the top-l subsets without making calls to the LLM.

Intuitively, given a subset S, the scoring function $\sigma(S)$ should model the validation loss of an exemplar subset, $L(\pi, \{S\}, \mathcal{V})$, since they are expected to generate identical rankings. Hence, we propose that σ should incorporate the relationship between the exemplar's question x_i and the validation example's question $u_j \in \mathcal{V}$. In this work, we capture the relationship using a similarity score, $E_{ij} = \frac{\phi(x_i)^T \phi(u_j)}{\|\phi(x_i)\|\|\phi(u_j)\|}$, where $\phi(x)$ is the feature encoding of a smaller transformer model, e.g. BERT. We model score as linear function of the similarity features E_{ij} of exemplars x_i in the subset S:

$$\sigma(\vec{\alpha}, S) = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} \alpha_i \mathbb{1}(x_i \in S) E_{ij} \quad (5)$$

Here, α_i denotes the i-th exemplar's contribution to the scoring function. We dynamically estimate α_i to fit the top-l (lowest loss) exemplar subsets in $\mathcal{S}$. Section 3.3 describes the details of the algorithm.

Note that since α_i's can be negative, they can be learned to implicitly estimate both positive and negative correlations between training exemplars x_i and x_j, according to signs of α_i and α_j. Also, note the above equation can be written as $\sum_{i=1}^n \alpha_i \mathbb{1}(x_i \in S) e_i$, where $e_i = \frac{1}{m}\sum_{j=1}^m E_{ij}$, which indicates that only aggregate effect of exemplars in validation set is modeled by σ. Finally, training exemplars in ICL can be thought of as representing different "distinguishable" solution "concepts" of tasks (see e.g. (Xie et al., 2021)). For instance, consider the exemplars selected by EXPLORA for the AQUARAT dataset (Table 7) can be identified with concepts like *Proportionality*, *Series*, *Kinematics*, and *Interest*. Since the final scoring function is trained to fit the top-l exemplar subsets resulting in low loss, the resultant α_i's can be thought of as implicitly representing the important concepts needed for effectively predicting the correct label (hence resulting in low loss). Unfortunately, since α_i's can be both positive and negative, their magnitudes are not directly interpretable as importance scores for the concepts. Unlike existing works e.g. (Li and Qiu, 2023; Xu and Zhang, 2024), we do not use probability scores for tokens provided by LLMs in our scoring function, as these scores are not directly related to the predictive task.

3.3 A Sampling-based bandit algorithm for Top-l Exemplar Subset Selection

Our exemplar-subset selection problem (Eq 4) in the context of the score function σ encompasses

Algorithm 1: EXPLORA

1 **Input:** $\mathcal{U} \subseteq \mathcal{S}$: ▷ Initial exemplar subsets
2 **Initialize:** $U_0 \leftarrow$ set of random l subsets from $\mathcal{U}$
3 $t \leftarrow 0$
4 $\vec{\alpha} \leftarrow \mathcal{N}(0,1)$ ▷ Sampling from a gaussian
5 **while** $t < T$ **do**
6 Let $V_t \leftarrow \mathcal{U} \setminus U_t$
7 $\vec{\alpha_t} \leftarrow \min_{\vec{\alpha}} \mathcal{L}(\vec{\alpha}, U_t, V_t)$ ▷ Eq. 6
8 $S_t^* = \arg\min_{S \in V_t} \sigma(\vec{\alpha_t}, S)$ ▷ Lowest loss subset
9 $\tilde{S}_t = \arg\max_{S \in U_t} \sigma(\vec{\alpha_t}, S)$ ▷ Highest loss subset
10 **if** $\sigma(\vec{\alpha_t}, S_t^*) < \sigma(\vec{\alpha_t}, \tilde{S}_t)$ **then**
11 $U_t \leftarrow U_t \setminus \{\tilde{S}_t\}$ ▷ Remove $\tilde{S}_t$
12 $U_{t+1} \leftarrow U_t \cup \{S_t^*\}$ ▷ add S_t^*
13 **end**
14 $t \leftarrow t + 1$
15 **end**
16 **Output:** U_T;Set of l subsets from $\mathcal{U}$ which have the lowest validation loss

two objectives: (1) learning the parameters of the loss model (σ) for the top-l exemplar subsets, and (2) calculating a set of low-loss exemplar subsets, U. This problem can be posed as the top-l arm selection problem in stochastic linear bandits (Kalyanakrishnan et al., 2012; Chaudhuri and Kalyanakrishnan, 2019). In the current setting, the arms correspond to the exemplar subsets S, and feature vectors for the linear bandit are given by E_{ij} where $x_i \in S$. The reward in our setting can be thought of as the negative of loss of a subset: $-L(\pi, \{S\}, \mathcal{V})$. *LUCB* (Kalyanakrishnan et al., 2012; Chaudhuri and Kalyanakrishnan, 2019) and later generalized variants of GIFA (Réda et al., 2021) are widely used for top-l arm selection in stochastic linear bandits. *LUCB* maintains two sets: (1) l high-reward arms (called U here), and (2) the other low-reward arms. In each round, one arm is pulled from each of the sets, leading to a revised estimate of the rewards of all the arms. However, these algorithms are impractical for our setting, since they require at least linear time in the number of arms, in each round.

Algorithm 1 describes EXPLORA, a novel sampling-based bandit algorithm, inspired by LUCB, for estimating α_i and identifying a set of l low-loss (corresponding to high reward) exemplar subsets U. *For practicality*, we start with a manageable set $\mathcal{U} \subseteq \mathcal{S}$ of exemplar subsets after eliminating the obvious ones (see Section 4). We initialize U_0 as l random subsets from $\mathcal{U}$. U_t in round t denotes the set of l subsets with the lowest loss. V_t denotes the set of other subsets (note that we do not exhaustively enumerate V_t). EXPLORA has two broad steps: (1) calculation of α_i based on losses computed using subsets from U_t and V_t, and (2) updating U_t based on the modified score function σ (due to updation of α_i). α_i is updated by minimizing the following loss function:

$$\mathcal{L}(\vec{\alpha}; U_t, V_t) = \sum_{S \in U_t} (L(S, \mathcal{V}) - \sigma(\vec{\alpha}, S))^2 + \sum_{S' \sim V_t} (L(S', \mathcal{V}) - \sigma(\vec{\alpha}, S'))^2 \quad (6)$$

Here the first term denotes the approximation error of the loss model for the low-loss set U_t, and the second term denotes the loss on *negative samples* from high-loss set V_t. The negative samples facilitate exploration over the set V_t by allowing the α_i values corresponding to unexplored and potentially low-loss subsets to be estimated correctly.

The key motivation behind this formulation is to be able to frugally compute α_i while making minimal calls to the LLM. A naive computation of the first term requires $l * m$ calls to the LLM. However, since U_{t+1} differs from U_t by only one exemplar subset, a caching mechanism can implement this step in m calls to the LLM, where m is the validation set size. The second term can be computed using $l' * m$ LLM calls, where l' is the number of negative samples. U_t is updated in lines 8 – 12 in Algorithm 1. U_{t+1} differs from U_t by only one exemplar subset. This leads to a smoother convergence of α_i over the iterations since the loss function $\mathcal{L}$ depends mainly on U_t. Line 11 removes the exemplar subset $\tilde{S}_t$ from U_t, which is the highest estimated loss subset in U_t, and line 12 adds S_t^* to U_t, which is the exemplar subset with the lowest estimated loss in V_t. While a formal convergence guarantee for the proposed algorithm will be explored elsewhere, the updates is designed to decrease the total validation loss of U_t, provided that the estimation of loss $\sigma(\vec{\alpha}, S), S \in U_t$ becomes more accurate over the iterations. This can be achieved by reducing l' over the rounds. Also, note that the step in line 8 can be expensive to implement in many settings, due to the size of V_t. Here one can perform an approximate search that finds a "good enough" S_t^* such that $\sigma(\vec{\alpha}, S_t^*) < \sigma(\vec{\alpha}, \tilde{S}_t)$.

4 Experimental Setup

We answer the following research questions.
RQ I. Can EXPLORA a static exemplar selection approach achieve competitive performance compared to the state-of-the-art?

Method	GSM8K	AquaRat	TabMWP	FinQA	StrategyQA
			GPT-3.5-turbo		
Dynamic					
KNN (Rubin et al., 2022)	53.45	51.96	78.33	51.52	81.83
KNN (S-BERT) (Rubin et al., 2022)	53.07	52.75	77.95	52.65	81.83
MMR (Ye et al., 2023b)	54.36	51.18	77.32	49.87	82.86
KNN+SC (Wang et al., 2023c)	80.21	62.59	83.08	54.49	83.88
MMR+SC (Wang et al., 2023c)	78.01	59.45	81.36	50.74	83.88
PromptPG (Lu et al., 2023b)	-	-	68.23	53.56	-
Static					
Zero-Shot COT (Kojima et al., 2023)	67.02	49.60	57.10	47.51	59.75
Manual Few-Shot COT (Wei et al., 2023)	73.46	44.88	71.22	52.22	73.06
Random	67.79	49.80	55.89	53.70	81.02
PS+ (Wang et al., 2023b)	59.30	46.00	-	-	-
Auto-COT (Zhang et al., 2023b)	57.10	41.70	-	-	71.20
GraphCut (Iyer and Bilmes, 2013)	66.19	47.24	60.45	52.31	80.00
FacilityLocation (Iyer and Bilmes, 2013)	68.61	48.43	67.66	36.79	81.63
LENS (Li and Qiu, 2023)	69.37	48.82	77.27	54.75	79.79
LENS+SC (Li and Qiu, 2023)	79.37	57.87	80.68	60.06	82.24
Our Approach					
EXPLORA	77.86(▲12.24%) †	53.54(▲9.67%)†	83.07(▲7.51%) †	59.46(▲8.60%) †	85.71(▲5.63%) †
EXPLORA+SC	**86.35**(▲24.48%) ‡	63.39(▲29.84%) ‡	85.52(▲10.68%) ‡	64.52(▲17.84%) ‡	87.14 (▲9.21%)†
EXPLORA+KNN+SC	85.14 (▲22.73%)‡	62.20(▲27.41%)‡	86.29(▲12.39%) ‡	**65.12**(▲18.94%) ‡	**88.37**(▲10.75%)†
EXPLORA+MMR+SC	86.13(▲24.16%) ‡	**63.78**(▲30.64%) ‡	**86.96**(▲12.54%)‡	64.60(▲17.99%) ‡	87.55(▲9.73%)†
			GPT-4o-mini		
LENS (Li and Qiu, 2023)	76.19	64.56	86.34	69.31	92.85
EXPLORA	93.63	69.29	90.12	72.71	95.10

Table 1: Results across datasets (we use 5-shot for all methods). Percentage improvements are reported over LENS (Li and Qiu, 2023). † indicates statistical significance using t-test over LENS at 0.05 level and ‡ at 0.01 level.

RQ II. Can we transfer the exemplars selected with respect to smaller language models directly to Larger Language Models?

RQ III. Can we minimize the number of calls to the language models during exemplar selection?

4.1 Experimental setting

Datasets, Metrics: We conduct extensive experiments over a range of complex reasoning datasets (GSM8K, AquaRat, TabMWP, FinQA and StrategyQA). We use official metrics of the datasets, i.e., *exact match* (EM), cover-EM (Press et al., 2023; Rosset et al., 2021). More details about the datasets and prompts can be found in Appendix A and C.

Hyperparameters: For all experiments, we set the temperature to 0.3 to mitigate randomness, with frequency and presence penalty set to 0.8 and 0.6 to avoid repetition. We set the max_token_length to 900 for generation. For efficiency reasons, we carry our experiments in a transfer setting, where we select exemplars using EXPLORA or other static exemplar selection methods for smaller models like Mistral-7b and Llama2-7b and then transfer them to a larger model like gpt-3.5-turbo to perform inference owing to its superior capabilities. We report performance on smaller LLMs in Appendix B.

Subset selection hyperparameters: For EXPLORA, we set k as 5, the desired number of clusters to be formed from the training set. We set the number of validation examples $\mathcal{V}$ to 20. We construct a set $\mathcal{U}$, of 40 subsets, each containing 5 exemplars, by randomly selecting one exemplar from each cluster with replacement. While we experimented with larger values for size of $\mathcal{U}$ (100..etc.) we observe that at $\mathcal{U}$=40, we achieve optimal performance on validation set. Initially, set U consists of 10 random subsets from set $\mathcal{U}$, while V comprises the remaining subsets not included in U. In each round, we randomly sample 5 subsets from V. We update U by removing the worst subset and add the best subset from V to it. This process repeats for 10 iterations, with the stopping criterion of approximation error being unchanged between iterations, resulting in U having 10 low-loss subsets.

EXPLORA Variants: We posit that static and dynamic approaches can complement each other and apply dynamic methods like MMR and KNN over the l subsets selected by EXPLORA, thereby reducing the search space. This makes EXPLORA+KNN and EXPLORA+MMR a hybrid approach.

4.2 Baselines

Exemplar selection : We compare with *dynamic exemplar selection methods* like similarity (KNN) (Rubin et al., 2022) and diversity (MMR) (Ye et al., 2023b). For KNN, we retrieve top 5 exemplars for each test example for a fair comparison with EXPLORA, and for MMR we observe $\lambda = 0.5$ to

Method	T	GSM	Aqua	Tab	Fin	Strat
EXP	L	79.07	53.94	80.11	54.66	85.31
	M	77.86	53.54	83.07	59.46	85.71
EXP+SC	L	85.82	63.78	86.76	61.16	85.10
	M	86.35	63.39	85.52	64.52	87.14
EXP+KNN+SC	L	85.89	64.17	85.74	63.64	86.53
	M	85.14	62.20	86.29	65.12	88.37
EXP+MMR+SC	L	86.20	62.99	87.81	64.60	86.12
	M	86.13	63.78	86.96	64.60	87.55

Table 2: Results for transfer (T) of exemplars selected using EXPLORA (EXP) on smaller LLMs (Llama2-7b (L) and Mistral-7b (M)) to larger LLM (gpt-3.5-turbo).

Datasets	GSM	Aqua	Tab	Fin	Strat
Zero-Shot COT	±5.18	±7.08	±1.84	±4.50	±4.19
Few-Shot COT	±4.48	±12.03	±1.66	±4.76	±5.67
KNN	±3.76	±5.49	±1.27	±4.17	±4.85
MMR	±4.00	±10.53	±1.68	±6.10	±5.70
Graph Cut	±6.38	±8.18	±2.03	±5.29	±7.62
Facility Location	±4.23	±6.71	±1.74	±4.94	±5.93
LENS	±5.04	±6.67	±1.59	±5.81	±3.98
EXPLORA	**±3.39**	**±4.93**	±1.50	**±3.41**	**±3.95**

Table 3: Comparison of robustness of EXPLORA to other approaches. We report standard deviation (lower is better) with scores from different splits of eval. set.

be the optimal value. For KNN (S-BERT), we employ sentence transformer *paraphrase-MiniLM-L6-v2*. We also compare with the *chain of thought methods* like Manual Few-Shot COT (Wei et al., 2023), Zero-Shot COT (Kojima et al., 2023), *random* , *coreset selection methods* (Facility Location and Graph Cut (Iyer and Bilmes, 2013)) and task-specific approaches like "Plan and Solve" (Wang et al., 2023b) and Auto-COT (Zhang et al., 2023b).

LENS (Li and Qiu, 2023): We compare with a closely related static exemplar selection method where the training data is filtered in two stages to extract informative examples.

5 Results

5.1 Performance Comparison

To answer **RQ1**, we compare EXPLORA and it's variants with state-of-the-art static and dynamic exemplar selection methods. We observe in Table 1 that EXPLORA outperforms the random baseline, which highlights the non-triviality of the proposed task of selecting task-level representative exemplars. We also observe that EXPLORA outperforms manual Few-Shot COT (Wei et al., 2023).

We also compare EXPLORA with LENS (Li and Qiu, 2023) a static exemplar selection method for ICL. We observe that EXPLORA and its variants significantly outperform LENS. For instance, on GSM8K EXPLORA outperforms LENS by **12.24%**. LENS scores each exemplar independently without considering any interactions between the exemplars, and also assumes access to LLM logits. However, in EXPLORA, scores are assigned to subsets, allowing for the implicit capture of the interplay between the exemplars within each subset. This is particularly important for reasoning tasks, as the exemplars need to contain sufficient information for solving diverse reasoning based questions. We perform a *qualitative analysis* of exemplars chosen by LENS vs EXPLORA in **Appendix D**.

We also observe that existing coreset selection methods like Graph Cut and Facility Location perform worse than or are similar in performance to the random exemplar selection. This indicates the importance of designing methods specific to ICL for exemplar selection.

We also observe that EXPLORA outperforms dynamic exemplar selection methods like KNN, PromptPG and MMR. A significant limitation, apart from additional inference time computational costs, is that dynamic exemplar selection methods do not consider interactions between the exemplars.

5.2 Transferability of exemplars from smaller LLMs to larger LLMs

To answer **RQ2**, we report the performance on test set in transfer setting across tasks using gpt-3.5-turbo with exemplars selected from Llama2-7b and Mistral-7b as shown in Table 2. In Table 1 we report the EXPLORA results from this transfer setup with exemplars selected from Mistral-7b. We observe that the exemplars selected by EXPLORA using smaller LLMs transfer well to larger LLMs, as indicated by their superior performance compared to baselines through evaluation in the transfer setting. This shows the strong transferability of our selected exemplars and the effectiveness of EXPLORA, which is robust across different LLMs. We attribute the transferability of the selected exemplars to design choice of EXPLORA, which remains agnostic to confidence scores from the LLMs.

5.3 Robustness of exemplars selected by EXPLORA compared to other approaches

We compare the robustness of EXPLORA to other exemplar selection methods. We measure standard deviation of performance across different subsets of the evaluation set through 10-fold cross validation, as shown in Table 3. We observe that the exemplars chosen by EXPLORA results in less variance across

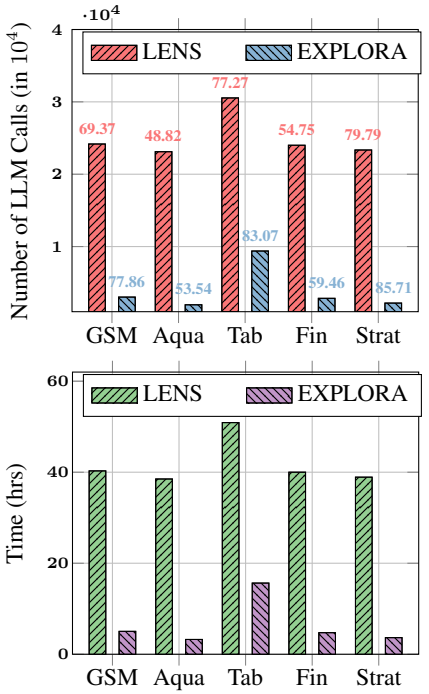

Figure 2: (Top) Frugal exemplar selection by EXPLORA: **LLM calls** LENS vs EXPLORA (y-axis) with corresponding EM scores indicated on top of bars. (Bottom) **Runtime** comparison LENS vs EXPLORA.

different subsets of the evaluation set when compared to other static exemplar selection methods. We also observe that in 3 out of 4 datasets, exemplars chosen by EXPLORA has less variance in task performance when compared to dynamic exemplar selection methods like KNN and MMR. Exemplars selected through dynamic approaches are not optimized for the task but rather on a per-test-example basis. Consequently, this leads to greater variance in final task performance. In TabMWP, we observe that the variance in results is low for all exemplar selection methods. Hence, EXPLORA helps select exemplars for the task which are more robust than other static methods or dynamic selection methods.

5.4 Exemplar selection efficiency based on number of LLM evaluations

To answer **RQ3**, we compare the number of calls made to the LLMs during the exemplar selection step. We compare the number of such calls and also running times for the LENS approach and our proposed approach EXPLORA as shown in Figure 2. We observe that LENS has significantly more LLM calls than EXPLORA (about **10x**). This is because LENS relies on confidence estimates from the LLM for each training example and computes an informativeness score for all examples in the dataset, incurring expensive LLM calls for each example. Whereas, EXPLORA computes scores for whole exemplar subsets and employs a exploration based approach, resulting in LLM calls only for small number of subsets. In summary, EXPLORA drastically reduces the number of LLM calls (to $\sim 11\%$ of calls made by LENS) and also reduces running time as shown in Figure 2 during exemplar selection step with significant performance gains.

Datasets	GSM	Aqua	Tab	Fin	Strat
Exhaustive eval	76.72	50.39	82.24	57.02	82.45
EXPLORA (-exploration) (Mistral)	75.89	50.00	75.16	50.30	80.40
EXPLORA (Llama)	**79.07**	**53.94**	80.11	54.66	**85.31**
EXPLORA (Mistral)	**77.86**	**53.54**	**83.07**	**59.46**	**85.71**

Table 4: Ablation studies: exhaustive evaluation, w/o exploration vs proposed exploration (EXPLORA).

5.5 Ablation Studies

We conduct an **exhaustive evaluation** to compare our exploration method and demonstrate its effectiveness, as shown in Table 4. Evaluating all possible subsets is infeasible, so we exhaustively evaluated a downsampled set of 40 subsets ($\mathcal{U}$ from EXPLORA) on the validation set. We then selected the subset with the minimum validation loss for subsequent inference on the test set. Our superior performance compared to the exhaustive evaluation is due to modeling the top-l subsets in each round.

Additionally, we perform an ablation study where we fit the linear model once on the entire downsampled set of 40 subsets (shown in Table 4). Then we select the subset with the minimum approximation error (loss) for subsequent inference on the test set. Note that in this ablation, the linear model is fit once for all subsets, whereas in EXPLORA, we model the top-l subsets in each round.

6 Conclusion

In this work, we propose an efficient and robust task level exemplar subset selection method that identifies highly informative exemplar subsets. The proposed method saves resources by reducing the number of LLM calls, in contrast to the current state-of-the-art. We also observe that the exemplars obtained using smaller LLMs can be well transferred to larger LLMs. EXPLORA outperforms existing static and dynamic exemplar selection methods. In future, we plan to further explore hybrid exemplar selection and the impact of exemplars for tasks involving complex reasoning.

7 Acknowledgements

We thank the reviewers and the meta-reviewer for their valuable and insightful feedback. We also thank the SERB grant "Online Subset Selection Algorithms for Data-centric Responsible AI and Efficient Computer Vision", sanction no. : CRG/2023/004600 Dt. 18-02-202 for supporting this project.

8 Contributions

Kiran contributed to experimental design, idea conceptualization, execution, **core algorithm (EXPLORA) implementation** and writing. Venktesh contributed to experimental design, idea conceptualization, execution, writing. Raghuram and Krishna ran experiments on smaller LLMs. Sourangshu contributed to formulation and writing of Section 3. Sourangshu and Avishek mentored the project and contributed to idea conceptualization and writing.

9 Limitations

Our method deals with selecting top-l exemplar subsets that are best suited to improve the overall performance through In-Context Learning (ICL). We identify certain limitations that could be addressed in future works.

Our method is significantly efficient and computationally less resource intensive compared to state-of-the-art exemplar selection methods. However, one of the limitations of our approach is that if the space of the subsets $\mathcal{U}$ is large in some scenarios, then the computational time of the step (Algo 1, line 8) that calculates the lowest loss subset from V would increase. However, it would not increase the number of LLM calls and would still be computationally less resource intensive than the existing approaches. We would also need more efficient ways to sample negative examples (Eq 6) with increase in size of $\mathcal{U}$. While currently we propose a random sampling mechanism to sample negative examples from V, in future we plan to further analyze the impact of sampling negative examples for larger $\mathcal{U}$ sizes. We defer this for future work, as it is beyond the scope of the current work.

While EXPLORA converges as observed from our experiments, the current work does not provide an analysis or provable guarantees for convergence of the parameter α. In future, we plan to provide an analysis for convergence of parameter α.

10 Ethical Considerations

The intended use of the proposed approach is exemplar selection for reasoning problems that can be used to build QA systems for finance or education. Since our approach uses LLM for complex reasoning-based QA, the risks of hallucination (Ji et al., 2023) must be taken into consideration before deploying the approach. Since users may trust the hallucinated answers from the QA system, this may result in the spread of misinformation (Zhang et al., 2023a; Albrecht et al., 2022). We observe that EXPLORA is more robust across test instances compared to baselines due to the transfer of informative exemplars with rationales. Although hallucination is still a possibility when employing EXPLORA and the resulting QA systems are not infallible.

Additionally, we do not use any private information for the proposed approach. Though LLMs may have been pre-trained on sensitive information, our prompts do not elicit any sensitive information directly or indirectly.

References

Joshua Albrecht, Ellie Kitanidis, and Abraham J. Fetterman. 2022. Despite "super-human" performance, current llms are unsuited for decisions about ethics and safety.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. 2019. Pac identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 991–1000. PMLR.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022b. Finqa: A dataset of numerical reasoning over financial data.

Ting-Rui Chiang and Yun-Nung Chen. 2019. Semantically-aligned equation generation for solving and reasoning math word problems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2656–2668, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021b. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. Deepcore: A comprehensive library for coreset selection in deep learning.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rishabh K Iyer and Jeff A Bilmes. 2013. Submodular optimization with submodular cover and submodular knapsack constraints. *Advances in neural information processing systems*, 26.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. 2012. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023a. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. 2021. Top-m identification for linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1108–1116. PMLR.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2021. Knowledge-aware language model pretraining.

Rishiraj Saha Roy and Avishek Anand. 2022. Question answering for the curated web: Tasks and methods in qa over knowledge bases and text collections.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems.

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

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners.

Jonathan Tonglet, Manon Reusens, Philipp Borchert, and Bart Baesens. 2023. Seer : A knapsack approach to exemplar selection for in-context hybridqa.

Venktesh V, Sourangshu Bhattacharya, and Avishek Anand. 2023. In-context ability transfer for question decomposition in complex qa.

V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 650–660. Association for Computing Machinery (ACM).

Dingzirui Wang, Longxu Dou, and Wanxiang Che. 2023a. A survey on table-and-text hybridqa: Concepts, methods, challenges and future directions.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxing Cao, Haiming Wang, Xiongwei Han, et al. 2023. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. *arXiv preprint arXiv:2310.02954*.

Shangqing Xu and Chao Zhang. 2024. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301*.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2022. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Jianyi Zhang, Xu Ji, Zhangchi Zhao, Xiali Hei, and Kim-Kwang Raymond Choo. 2023a. Ethical considerations and policy implications for large language models: Guiding responsible development and deployment.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 418–426.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

A Datasets Description

An overview of the dataset statistics and examples are shown in Table 5.

FinQA: Comprises financial questions over financial reports that require numerical reasoning with structured and unstructured evidence. Here, 23.42% of the questions only require the information in the text to answer; 62.43% of the questions only require the information in the table to answer; and 14.15% need both the text and table to answer. Meanwhile, 46.30% of the examples have one sentence or one table row as the fact; 42.63% has two pieces of facts; and 11.07% has more than two pieces of facts. This dataset has 1147 questions in the evaluation set.

AquaRat: This dataset comprises 100,000 algebraic word problems in the train set with dev and test set each comprising 254 problems. the problems are provided along with answers and natural language rationales providing the step-by-step solution to the problem. An examples problem is shown in Table 5.

TabMWP: It is a tabular-based math word problem-solving dataset with 38,431 questions. TabMWP is rich in diversity, where 74.7% of the questions in TabMWP belong to free-text questions, while 25.3% are multi-choice. We treat all questions as free-form type and do not provide any options to the LLM for consistent evaluation. We evaluate on the test set with 7686 problems.

GSM8K: This dataset consists of linguistically diverse math problems that require multi-step reasoning. The dataset consists of 8.5K problems and we evaluate on the test set of 1319 questions.

StrategyQA: To prove the generality of our approach for reasoning tasks, we evaluate on StrategyQA (Geva et al., 2021b), a dataset with implicit and commonsense reasoning questions. Since there is no public test set with ground truth answers, we perform stratified sampling done on 2290 full train set to split into 1800 train and 490 test.

Metrics: For TabMWP and StrategyQA we employ cover-EM (Rosset et al., 2021; Press et al., 2023), a relaxation of Exact Match metric which checks whether the ground truth answer is contained in the generated answer. This helps handle scenarios where LLM generates "4 hours" and the ground truth is "4". For other numerical reasoning datasets, we employ Exact match.

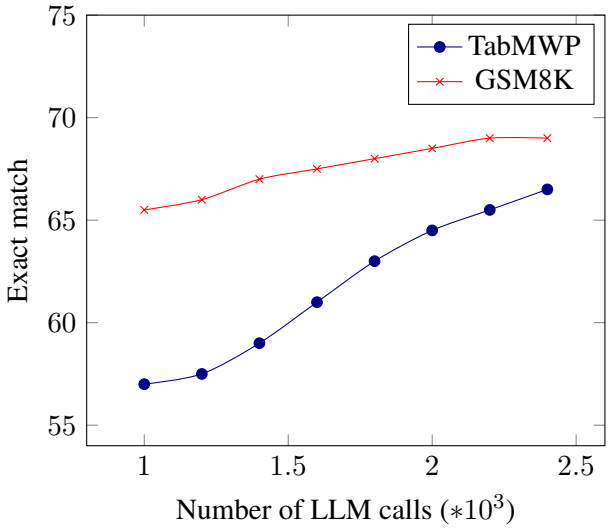

Figure 3: Plot showing number of LLM calls vs Exact Match for TabMWP and GSM8K

B Results using Alternate Open Source LLMs

We also report the performance of the proposed exemplar selection approach EXPLORA on open-source models like Mistral-7b and LLama2-7b. The results are shown in Table 6. We observe that the absolute performance across baselines and EXPLORA is lower than when employing gpt-3.5-turbo as backbone for the same exemplars. We primarily observe that this is due to the scale of the Language models as Mistral and LLAMA2 models have 7 billion parameters while gpt-3.5-turbo is of much larger scale and the emergent capabilities like In-Context Learning are proportional to scale of the language models (Wei et al., 2022).

However, we still observe that EXPLORA leads to reasonable performance gains over other static exemplar selection methods across the smaller open-source LLMs. We also observe that EXPLORA and its variants are competitive with dynamic exemplar selection methods.

Our main experiments are carried out in a transfer setting where exemplar selection is done using small open source LLMs and transferred to larger LLMs. This is done for reducing the cost of LLM inference during exemplar selection, and also to leverage superior performance of LLMs with larger scale during inference. This setting is inspired from the work μP (Yang et al., 2022) where the language model hyperparameters are tuned on a smaller LM and transferred to a larger language model.

C Prompts

We also demonstrate the instructions issued to the LLM for different tasks discussed in this work, along with some exemplars selected using EXPLORA. An example of prompt construction for FinQA is shown in Figure 5. We also showcase example prompts for AquaRat (Figure 4), GSM8K (Figure 6), TabMWP (Figure 7) and StrategyQA (Figure 8).

D Exemplar Qualitative Analysis

We provide a qualitative analysis of exemplars and compare the exemplars selected using EXPLORA with exemplars selected using LENS (Li and Qiu, 2023), the recent state-of-the-art approach.

The final set of exemplars chosen by LENS vs EXPLORA for the AquaRat dataset is shown in Table 7. We observe that Question 4 and Question 5 in the set of exemplars chosen by LENS are redundant in that they are very similar problems that require similar reasoning steps and are also similar thematically. Both the questions are centered on the theme of work and time and are phrased in a similar manner. Hence, they do not add any additional information to solve diverse problems the LLM may encounter during inference. However, we observe that the exemplars chosen by EXPLORA are problems that require diverse reasoning capabilities and are also different thematically.

We also compare the exemplars chosen by EXPLORA with LENS for the FinQA dataset (Table 8) and make similar observations. We observe that the exemplars chosen by EXPLORA comprises diverse set of problems with diverse reasoning. We also observe that EXPLORA also contains exemplars that require composite numerical operations with multi-step reasoning rationales to arrive at the solutions, whereas LENS mostly has exemplars with single-step solutions.

The exemplars chosen by LENS compared to EXPLORA for TabMWP are shown in Table 10. We observe that exemplar 1 and exemplar 3 chosen by LENS are redundant, as they represent the same reasoning concept of computing median for a list of numbers. However, we observe that EXPLORA selects diverse exemplars with each exemplar representing a different reasoning concept.

We also demonstrate the exemplars for GSM8K and StrategyQA in Table 9 and Table 11 respectively.

E Analysis of Accuracy (Exact Match) vs number of LLM calls

We analyze the change in accuracy in proportion to number of calls to the LLM (EXPLORA Algo 1 iterations) as shown in Figure 3. We observe that the performance increases with number of LLM calls/iterations (Algorithm 1) of EXPLORA algorithm. We also observe that for GSM8K and TabMWP EXPLORA converges and obtains optimal validation set performance quickly with less number of LLM calls, as observed in Figure 3.

Dataset	#Train	#Test	Example Question	Description
GSM8K (Cobbe et al., 2021)	7473	1319	`Claire makes a 3 egg omelet every morning for breakfast. How many dozens of eggs will she eat in 4 weeks?`	multi-step arithmetic word problems
AquaRat (Ling et al., 2017)	97467	254	`A trader sold an article at a profit of 20% for Rs.360. What is the cost price of the article?`	multi-step arithmetic word problems
TabMWP (Lu et al., 2023a)	23059	7686	`Allie kept track of how many kilometers she walked during the past 5 days. What is the range of the numbers?`	Table based numerical reasoning
FinQA (Chen et al., 2022b)	6251	1147	`In 2010 and 2009, what was total fair value in billions of assets segregated for benefit of securities and futures brokerage customers?`	Table and Text based numerical reasoning
StrategyQA (Geva et al., 2021a)	1800	490	`Would the chef at Carmine's restaurant panic if there was no basil?`	multi-step reasoning

Table 5: Overview of the Complex QA datasets used in this study.

Method	GSM8K	AquaRat	TabMWP	FinQA
		Mistral-7B		
dynamic				
KNN (Rubin et al., 2022)	37.98	23.22	61.74	9.06
MMR (Ye et al., 2023b)	46.25	18.11	52.82	10.11
static				
Zero-Shot COT (Kojima et al., 2023)	7.43	21.65	43.34	1.74
Manual Few-shot COT (Wei et al., 2023)	30.48	14.90	46.94	3.22
Random	33.81	22.04	30.40	5.52
GraphCut (Iyer and Bilmes, 2013)	47.00	21.65	59.46	5.66
FacilityLocation (Iyer and Bilmes, 2013)	46.25	14.17	57.74	4.96
LENS (Li and Qiu, 2023)	46.39	29.92	57.69	5.14
Our Approach				
EXPLORA	46.41	30.07	58.62	5.57
EXPLORA+SC	**61.41**	**35.10**	**61.96**	7.41
		Llama2-7B		
dynamic				
KNN (Rubin et al., 2022)	23.43	28.34	54.83	10.37
MMR (Ye et al., 2023b)	29.64	21.65	49.61	12.20
static				
Zero-Shot COT (Kojima et al., 2023)	6.14	6.29	40.31	1.67
Manual Few-shot COT (Wei et al., 2023)	21.15	20.47	43.23	2.87
Random	18.27	21.60	28.41	4.62
GraphCut (Iyer and Bilmes, 2013)	27.29	21.65	47.38	5.40
FacilityLocation (Iyer and Bilmes, 2013)	27.21	21.65	47.23	5.05
LENS (Li and Qiu, 2023)	28.05	23.22	48.29	6.62
Our Approach				
EXPLORA	28.67	24.40	49.96	6.62
EXPLORA+SC	**36.85**	24.01	**56.74**	6.21

Table 6: Results across datasets on MISTRAL-7B and LLAMA-2-7B.

AQUA Prompt

Instruction:`You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math. Follow given examples and solve the problems in step by step manner.`

Exemplars :
[Question]: *The average age of three boys is 45 years and their ages are in proportion 3:5:7. What is the age in years of the youngest boy?*
[Options]: A) 9, B) 10, C) 11, D) 12, E) 13
[Explanation]: $3x + 5x + 7x = 45$,
$x = 3$,
$3x = 9$
[Answer]: The option is A
...
...

Test Input : Question: Options:
Explanation: [INS] Answer: [INS]

Figure 4: Prompt for AquaRat

FinQA Prompt

Instruction:`You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math, using the information from given table and text.`

Exemplars :
Read the following table, and then answer the question:
[Table]: Year | 2016 | 2015 | 2014 |
share-based compensation expense | 30809 | 21056 | 29793 |
income tax benefit | 9879 | 6907 | 7126 |
[Question]: *how much percent did the income tax benefit increase from 2014 to 2016?*
[Explanation]: $x0 = (9879 - 7126)$,
ans=($x0/7126$)
[Answer]: The answer is increased 38.6%
...
...

Test Input : Read the following table, and then answer the question: Table: Question:
Explanation: [INS] Answer: [INS]

Figure 5: Prompt for FinQA

GSM8K Prompt

Instruction:`You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math. Follow given examples and solve the problems in step by step manner.`

Exemplars :
[Question]: *Samir just turned half the age Hania was* 10 *years ago. If in five years Hania will be* 45 *years old, what will Samir's age be five years from now?*
[Explanation]: If in five years, Hania will be 45 years old, currently she is $45 - 5 = 40$ years old.
Samir just turned half the age Hania was 10 years ago, which means she is $30/2 = 15$ years old.
In five years, Samir will be $15 + 5 = 20$ years old.
[Answer]: 20 years old
...
...

Test Input : Question:
Explanation: [INS] Answer: [INS]

Figure 6: Prompt for GSM8K

TabMWP Prompt

Instruction:`You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math, using the information from the given table. Solve the given problem step by step providing an explanation for your answer.`

Exemplars :
[Table]: Table: Day | Number of tickets
Monday | 36
Tuesday | 43
Wednesday | 46
Thursday | 59
Friday | 37
Saturday | 46
Sunday | 51
[Question]: *The transportation company tracked the number of train tickets sold in the past 7 days. What is the range of the numbers?*
[Explanation]: Read the numbers from the table. 36, 43, 46, 59, 37, 46, 51
First, find the greatest number. The greatest number is 59.
Next, find the least number. The least number is 36.
Subtract the least number from the greatest number: 59 - 36 = 23
[Answer]: The range is 23
...
...

Test Input : Table: Question:
Explanation: [INS] Answer: [INS]

Figure 7: Prompt for TabMWP

StrategyQA Prompt

Instruction:`You are a helpful, respectful and honest assistant helping to solve commonsense problems requiring reasoning. Follow the given examples that use the facts to answer a question by decomposing into sub-questions first and then predicting the final answer as "Yes" or "No" only.`

Exemplars :
[Facts]: Snowden scored above 145 on two separate IQ tests. The minimum accepted IQ score for MENSA on the Stanford–Binet is 132, while for the Cattell it is 148.
[Question]: *Could Edward Snowden join MENSA?*
[Sub-question 1]: What is the minimum accepted IQ score to be admitted to MENSA?
[Sub-question 2]: What is Edward Snowden's IQ?
[Sub-question 3]: Is #2 greater than or equal to #1?
[Answer]: Yes
...
...

Test Input : Facts: Question:
Sub-question: [INS] Answer: [INS]

Figure 8: Prompt for StrategyQA

Method	Exemplars
LENS	**Question**: A cat chases a rat 6 hours after the rat runs. cat takes 4 hours to reach the rat. If the average speed of the cat is 90 kmph, what s the average speed of the rat? **Options**: ['A)32kmph', 'B)26kmph', 'C)35kmph', 'D)36kmph', 'E)32kmph'] **Rationale:** Cat take 10 hours and rat take 4 hours...then Distance is 90*4.so speed of rat is (90*4)/10 = 36kmph **Answer: D**
	Question: A business executive and his client are charging their dinner tab on the executive's expense account.The . . . ? **Options**: ['A)69.55$', 'B)50.63$', 'C)60.95$', 'D)52.15$', 'E)53.15'] **Rationale**: let x is the cost of the food 1.07x is the gross bill after including sales tax 1.15* 1.07x=75 **Answer**: C
	Question: John and David were each given X dollars in advance for each day they were expected to perform at a community festival. John eventually,. . . ? **Options**: 'A)11Y', 'B)15Y', 'C)13Y', 'D)10Y', 'E)5Y' **Rationale:** . . . **Answer:** A
	Question:A contractor undertakes to do a piece of work in 40 days. He engages 100 men at the beginning and 100 more after 35 days and completes the work in stipulated time. If he had not engaged the additional men, how many days behind schedule would it be finished?? **Options**: 'A)2', 'B)5', 'C)6', 'D)8', 'E)9' **Rationale:** [(100$\times$ 35)+(200$\times$ 5)] men can finish the work in 1 day therefore 4500 men can finish the work in 1 day. 100 men can finish it in $\frac{4500}{100}$ = 45 days. This is 5 days behind Schedule **Answer:** A
	Question: A can do a job in 9 days and B can do it in 27 days. A and B working together will finish twice the amount of work in ——- days? **Options**: 'A)22 days', 'B)18 days', 'C)22 6/2 days', 'D)27 days', 'E)9 days' **Rationale:** 1/9 + 1/27= 3/27 = 1/9 9/1 = 9*2 =18 day **Answer:** B
EXPLORA	**Question**: The average age of three boys is 15 years and their ages are in proportion 3:5:7. What is the age in years of the youngest boy? **Options**: ['A)9', 'B)10', 'C)11', 'D)12', 'E)13'] **Rationale:** 3x + 5x + 7x = 45, x =3, 3x = 9 **Answer: A**
	Question: Can you deduce the pattern and find the next number in the series? $6, 14, 26, 98$? **Options**: ["A)276', 'B)277', 'C)278', 'D)279', 'E)None of these'] **Rationale**: $6 = 1^1 + 2^1 + 3^1, 14 = 1^2 + 2^2 + 3^2, 36 = 1^3 + 2^3 + 3^3, 98 = 1^4 + 2^4 + 3^4$ Thus the next number **Answer**: A
	Question: In covering a distance of 42 km, A takes 2 hours more than B. If A doubles his speed, then he would take 1 hour less than B. A's speed is:? **Options**: 'A)5 km/h', 'B)7 km/h', 'C)10 km/h', 'D)15 km/h', 'E)25 km/h' **Rationale:** Let A's speed be X km/hr. Then, 42/x - 42/2x = 3 6x = 42 x = 7 km/hr **Answer:** B
	Question: Find the number which when multiplied by 15 is increased by 196. **Options**: 'A)14', 'B)20', 'C)26', 'D)28', 'E)30' **Rationale:** Solution Let the number be x . Then, 15x - x = 196 ‹=›14x = 196 x ‹=› 14 **Answer:** A
	Question: A certain sum of money at simple interest amounted Rs.980 in 3 years at 5% per annum, find the sum? **Options**: 'A)867', 'B)855', 'C)299', 'D)852', 'E)903' **Rationale:** 980 = P [1 + (3*5)/100] P = 852 **Answer:** D

Table 7: Qualitative analysis of exemplars for **AquaRat** dataset selected by LENS vs EXPLORA. Rationale is not completely shown for some questions to conserve space. However, in our experiments all exemplars include rationales.

Method	Exemplars
LENS	**Table**: ǀ increase (decrease) ǀ average yield ǀ 2.75% (2.75 %) ǀ volume ǀ 0.0 to 0.25 ǀ energy services ǀ 2013 ǀ fuel recovery fees ǀ 0.25 ǀ recycling processing and commodity sales ǀ 0.25 to 0.5 ǀ acquisitions / divestitures net ǀ 1.0 ǀ total change ǀ 4.25 to 4.75% (4.75 %) ǀ **Question**: what is the ratio of the acquisitions / divestitures net to the fuel recovery fees as part of the expected 2019 revenue to increase **Rationale:** ans=(1.0 / 0.25) **Answer:** The answer is 4
	Table: (in millions) ǀ 2009 ǀ 2008 ǀ 2007 ǀ sales and transfers of oil and gas produced net of production andadministrative costs ǀ -4876 (4876) ǀ -6863 (6863) ǀ -4613 (4613) ǀ . . . **Question**: were total revisions of estimates greater than accretion of discounts? **Rationale**: . . . **Answer**: The answer is yes
	Table: ǀ 2007 ǀ 2008 ǀ change ǀ capital gain distributions received ǀ 22.1 ǀ 5.6 ǀ -16.5 (16.5) ǀ other than temporary impairments recognized ǀ -.3 (.3) ǀ -91.3 (91.3) ǀ -91.0 (91.0) ǀ net gains (losses) realized onfund dispositions ǀ 5.5 ǀ -4.5 (4.5) ǀ -10.0 (10.0) ǀ net gain (loss) . . . **Question**: what percentage of tangible book value is made up of cash and cash equivalents and mutual fund investment holdings at december 31 , 2009? **Rationale:** (1.4 / 2.2) **Answer:** The answer is 64%
	Table: in millions ǀ 2009 ǀ 2008 ǀ 2007 ǀ sales ǀ 5680 ǀ 6810 ǀ 6530 ǀ operating profit ǀ 1091 ǀ 474 ǀ 839 ǀ **Question**: north american printing papers net sales where what percent of total printing paper sales in 2009? **Rationale:** x0=(2.8 * 1000), ans=(x0 * 5680) **Answer:** The answer is 49%
	Table: in millions ǀ december 312015 ǀ december 312014 ǀ total consumer lending ǀ 1917 ǀ 2041 ǀ total commercial lending ǀ 434 ǀ 542 ǀ total tdrs ǀ 2351 ǀ 2583 ǀ nonperforming ǀ 1119 ǀ . . . **Question**: what was the change in specific reserves in all between december 31 , 2015 and december 31 , 2014 in billions? **Rationale:** (.3 - .4) **Answer:** The answer is -0.1
EXPLORA	**Table**: ǀ 2008 ǀ 2007 ǀ balance at beginning of year ǀ 23.2 ǀ 56.4 ǀ additions due to acquisition of allied ǀ 582.9 ǀ 2014 ǀ additions based on tax positions related to current year ǀ 10.6 ǀ 16.3 ǀ reductions for tax positions related to the current year ǀ -5.1 (5.1) ǀ -17.2 (17.2) ǀ . . . **Question**: in 2008 what was the change in the gross unrecognized tax benefits in millions **Rationale:** (611.9 - 23.2) **Answer:** The answer is 588.7
	Table: december 31 2004 ǀ 1054 ǀ december 31 2005 ǀ 1216 ǀ december 31 2006 ǀ 1219 ǀ december 31 2007 ǀ 2566 ǀ **Question**: what was devon's average translation adjustments included in accumulated other comprehensive income (in millions) from 2004 through 2007? **Rationale**: x0=(1054 + 1216), x1=(x0 + 1219), x2=(x1 + 2566), ans=(x2 + 4) **Answer**: The answer is 1513.75 **Table:** ǀ 2016 ǀ 2015 (in thousands) ǀ 2014 ǀ share-based compensation expense ǀ 30809 ǀ 21056 ǀ 29793 ǀ income tax benefit ǀ 9879 ǀ 6907 ǀ 7126 ǀ **Question**: how much percent did the income tax benefit increase from 2014 to 2016? **Rationale:** x0=(9879 - 7126), ans=(x0 - 7126) **Answer:** The answer is increased 38.6% **Table:** in billions ǀ 2018 ǀ january 1 ǀ 33.3 ǀ issuances ǀ 4.5 ǀ calls and maturities ǀ -6.8 (6.8) ǀ other ǀ -.1 (.1) ǀ december 31 ǀ 30.9 ǀ **Question**: assuming all matured securities were pledged as collateral , how much should we assume came from the calls? **Rationale:** ans=(6.8 - 4.9) **Answer:** The answer is 1.9 **Table:** in millions of dollars ǀ u.s . ǀ outside u.s . ǀ december 31 2008 ǀ december 31 2007 ǀ commercial and similar letters of credit ǀ 2187 ǀ 6028 ǀ 8215 ǀ 9175 ǀ . . . **Question**: what percentage of citigroup 2019s total other commitments as of december 31 , 2008 are outside the u.s.? **Rationale:** ans=(236931 / 1349500) **Answer:** The answer is 18%

Table 8: Qualitative analysis of exemplars for **FinQA** dataset selected by LENS vs EXPLORA. Rationale is not completely shown for some questions to conserve space. However, in our experiments all exemplars include rationales.

Method	Exemplars
LENS	**Question**: Michael wants to dig a hole 400 feet less deep than twice the depth of the hole that his father dug. The father dug a hole at a rate of 4 feet per hour. If the father took 400 hours to dig his hole . . . ? **Rationale:** Since the father dug a hole with a rate of 4 feet per hour, if the father took 400 hours digging the hole, he dug a hole 4*400 = 1600 feet deep. . . . Michael will have to work for 2800/4 = 700 hours. **Answer:** 700
	Question: When Erick went to the market to sell his fruits, he realized that the price of lemons had risen by 4 for each lemon. The price of grapes had also increased by half the price that . . . ? **Rationale**: The new price for each lemon after increasing by 4 is $8 + 4 = 12$ For the 80 lemons, . . . Erick collected $140 * 9 = 1260$ From the sale of all of his fruits, Erick received $1260 + 960 = 2220$. **Answer**: 2220
	Question: James decides to build a tin house by collecting 500 tins in a week. On the first day, he collects 50 tins. On the second day, he manages to collect 3 times that number. . . . ? **Rationale:** On the second day, he collected 3 times the number of tins he collected on the first day, which is $3 * 50 = 150$ tins. . . . he'll need to collect $200/4 = 50$ tins per day to reach his goal. **Answer:** 50
	Question: Darrel is an experienced tracker. He can tell a lot about an animal by the footprints it leaves behind. Based on the impressions, he could tell the animal was traveling east at 15 miles per hour . . . ? **Rationale:** If we let x be the amount of time, in hours, it will take for Darrel to catch up to the coyote, . . . If we subtract 1 x from each side, we get x=1, the amount of time in hours. **Answer:** 1
	Question: Martha needs to paint all four walls in her 12 foot by 16 foot kitchen, which has 10 foot high ceilings . . . If Martha can paint 40 square feet per hour, how many hours will it take her to paint kitchen? **Rationale:** There are two walls that are 12' by 10' and two walls that are 16' by 10' . . . how many hours she needs to finish: 1680 sq ft / 40 sq ft/hour = 42 hours **Answer:** 42
EXPLORA	**Question**: The difference between the number of boys and girls in a tree planting event is 400. If there are 600 boys at the event, and the number of girls is more than the number of boys . . . ? **Rationale:** If there are 600 boys at the event, and the difference between boys and girls is 400, there are 600+400=1000 girls. . . . 60% of the total number of boys and girls at the event is 60/100*1600=960 **Answer:** 960
	Question: Casey is trying to decide which employee she wants to hire. One employee works for $20 an hour. The other employee works for $22 an hour, but Casey would also get a $6/hour subsidy . . . ? **Rationale**: First find the weekly cost of the first employee: $20/hour * 40 hours/week = 800/week . . . subtract the smaller weekly cost from the bigger weekly cost: $800/week - $640/week = 160/week. **Answer**: 160 **Question:**Cara has 60 marbles in a bag. 20 of them are yellow, half as many are green, . . . If Cara picks a marble at random, what are the odds it's blue (expressed as a percentage)? **Rationale:** First find the number of green marbles: 20 yellow marbles/2=10 green marbles. . . . find the chance of getting a blue marble: 15/60 marbles*100=25%. **Answer:** 25 **Question:** Samir just turned half the age Hania was 10 years ago. If in five years Hania will be 45 years old, what will Samir's age be five years from now? **Rationale:** If in five years, Hania will be 45 years old, currently she is 45-5=40 years old. Ten years ago, Hania was 40-10=30 years old. . . . In five years, Samir will be 15+5=20 years old. **Answer:** 20
	Question: If the normal hours of operation of Jean's business are 4 pm to 10p every day Monday through Friday, and from 6 pm to 10 pm on weekends, how many hours is the business open in a week? **Rationale:** First, we find the number of hours per weekday by subtracting the smaller number from the larger one . . . We add these two totals together to find that 30+8=38 hours per week. **Answer:** 38

Table 9: Qualitative analysis of exemplars for **GSM8K** dataset selected by LENS vs EXPLORA. Rationale is not completely shown for some questions to conserve space. However, in our experiments all exemplars include rationales.

Method	Exemplars
LENS	**Table**: | Name | Age (years) | Jessica | 2 | Dalton | 7 | Kelsey | 5 | Lamar | 8 | Alexis | 2 **Question**: A girl compared the ages of her cousins. What is the median of the numbers? **Rationale:** Read the numbers: 2, 7, 5, 8, 2. Arrange the numbers in ascending: 2, 2, 5, 7, 8. Find the number in middle. The number in middle is 5. The median is 5. **Answer:** 5
	Table: | City | Number of houses sold | Melville | 878 | New Hamburg | 871 | Charles Falls | 881 | Pennytown | 817 **Question**: A real estate agent looked into how many houses were sold in different cities. Where were the fewest houses sold? **Rationale**: Find the least number in the table. . . . Pennytown corresponds to 817. **Answer**: 817
	Table: | Day | Number of new customers | Saturday | 2 | Sunday | 2 | Monday | 9 | Tuesday | 4 | Wednesday | 10 | Thursday | 3 | Friday | 6 **Question**: A cable company paid attention to how many new customers it had each day. What is the median of the numbers? **Rationale:** Find the number in middle. The number in middle is 4. The median is 4. **Answer:** 4
	Table: | Day | Number of cups | Friday | 8 | Saturday | 4 | Sunday | 10 | Monday | 6 | Tuesday | 6 | Wednesday | 1 | Thursday | 0 **Question**: Nancy wrote down how many cups of lemonade she sold in the past 7 days. What is the range of the numbers? **Rationale:** Subtract the least number from the greatest: 10-0=10. The range is 10. **Answer:** 10
	Table: | Price | Quantity demanded | Quantity supplied |$700 | 9,800 | 22,600 |$740 | 8,000 | 22,800 |$780 | 6,200 | 23,000 |$820 | 4,400 | 23,200 |$860 | 2,600 | 23,400 **Question**: At a price of $860, is there a shortage or a surplus? **Rationale:** At price of $860, quantity demanded is less than quantity supplied. . . . So, there is a surplus. **Answer:** surplus
EXPLORA	**Table**: | Day | Number of tickets | Monday | 36 | Tuesday | 43 | Wednesday | 46 | Thursday | 59 | Friday | 37 | Saturday | 46 | Sunday | 51 **Question**: The transportation company tracked the number of tickets sold in the past 7 days. What is the range of the numbers? **Rationale:** Find greatest number. . . . Subtract least from the greatest number: 59 - 36 = 23. The range is 23. **Answer:** 23 **Table:** | Stem | Leaf | 4 | 2, 7, 9, 9, 9 | 5 | 1 | 6 | 9 | 7 | 2, 2, 3, 3, 5 | 8 | | 9 | 0 **Question**: A pottery factory kept track of the number of broken plates per shipment last week. How many shipments had exactly 73 broken plates? **Rationale**: For the number 73, the stem is 7, and the leaf is 3. Find the row where the stem is 7. In that row, count all the leaves equal to 3. . . . 2 shipments had exactly 73 broken plates. **Answer**: 2
	Table: | purple and red clay bead | $0.02 | small pink bead | $0.04 | pearl bead | $0.07 | round silver bead | $0.01 | brown cat's eye bead | $0.08 | orange glass bead | $0.07 **Question**: Kylie has $0.05. Does she have enough to buy a small pink and purple and red clay bead? **Rationale:** Add the price of a small pink and purple and red clay bead: $0.04 + $0.02 = $0.06. $0.06 is more than $0.05. **Answer:** Kylie does not have enough money. **Table:** | Price | Quantity demanded | Quantity supplied | $165 | 17,900 | 6,400 | $345 | 15,100 | 8,900 | $525 | 12,300 | 11,400 | $705 | 9,500 | 13,900 | $885 | 6,700 | 16,400 **Question**: Look at the table. Then answer the question. At a price of $885, is there a shortage or a surplus? **Rationale:** At the price of $885, the quantity demanded is less than the quantity supplied. . . . So, there is a surplus. **Answer:** surplus **Table:** | Chickenville | 3:00 A.M. | 12:00 P.M. | 3:30 P.M. | Floral Gardens | 3:45 A.M. | 12:45 P.M. | 4:15 P.M. | Pleasant River Campground | 4:45 A.M. | 1:45 P.M. | 5:15 P.M. | Happy Cow Farm | 5:15 A.M. | 2:15 P.M. | 5:45 P.M. | Rocky Ravine Town | 5:45 A.M. | 2:45 P.M. | 6:15 P.M. **Question**: Look at the following schedule. Doug just missed the 3.00 A.M. train at Chickenville. How long does he have to wait until the next train? **Rationale:** Find 3:00 A. M. in the row for Chickenville. Look for the next train in that row. The next train is at 12:00 P. M. The elapsed time is 9 hours. **Answer:** 9

Table 10: Qualitative analysis of exemplars for **TabMWP** dataset selected by LENS vs EXPLORA. Rationale is not completely shown for some questions to conserve space. However, in our experiments all exemplars include rationales.

Method	Exemplars
LENS	**Facts**: Penguins are native to deep, cold parts of southern hemisphere. Miami is located in the northern hemisphere and has a warm climate. **Question:** Would it be common to find a penguin in Miami? **Rationale:** Where is a typical penguin's natural habitat? What conditions make #1 suitable for penguins? Are all of #2 present in Miami? **Answer: No**
	Facts: Shirley Bassey recorded the song Diamonds are Forever in 1971. Over time, diamonds degrade and turn into graphite. Graphite is the same chemical composition found in pencils. **Question:** Is the title of Shirley Bassey's 1971 diamond song a true statement? **Rationale**: What is the title to Shirley Bassey's 1971 diamond song? Do diamonds last for the period in #1? **Answer**: No
	Facts: The first six numbers in Fibonacci sequence are 1,1,2,3,5,8. Since 1 is doubled, there are only five different single digit numbers. **Question:** Are there five different single-digit Fibonacci numbers? **Rationale:** What are the single-digit numbers in Fibonacci sequence? How many unique numbers are in #1? Does #2 equal 5? **Answer:** Yes
	Facts: Katy Perry's gospel album sold about 200 copies. Katy Perry's most recent pop albums sold over 800,000 copies. **Question:** Do most fans follow Katy Perry for gospel music? **Rationale:** What type of music is Katy Perry known for? Is Gospel music the same as #1? **Answer:** No
	Facts: The Italian Renaissance was a period of history from the 13th century to 1600. A theocracy is a type of rule in which religious leaders have power. Friar Girolamo Savonarola was the ruler of Florence, after driving out the Medici family, from November 1494 â€ 23 May 1498. **Question:** Was Florence a Theocracy during Italian Renaissance? **Rationale:** When was the Italian Renaissance?When did Friar Girolamo Savonarola rule Florence? Is #2 within the span of #1? Did Friar Girolamo Savonarola belong to a religious order during #3? **Answer:** Yes
EXPLORA	**Facts**: The average cost of a US Boeing 737 plane is 1.6 million dollars. Wonder Woman (2017 film) grossed over 800 million dollars at the box office. **Question**: Is a Boeing 737 cost covered by Wonder Woman (2017 film) box office receipts? **Rationale:** How much does a Boeing 737 cost?. How much did the 2017 movie Wonder Woman gross? Is #2 greater than #1? **Answer:** Yes
	Facts: Big Show is a professional wrestler that weighs 383 pounds. Force is equal to mass times acceleration. An adult Cheetah weighs around 160 pounds. An adult Cheetah can run up to 58 MPH. **Question:** Can a cheetah generate enough force to topple Big Show? **Rationale**: How much does Big Show weigh? How much does a cheetah weigh? How fast can a cheetah run? Is the force produced by a mass of #2 and a speed of #3 enough to knock over something that weighs #1? **Answer**: Yes **Facts**: Spaghetti and meatballs are a staple on Italian pizzeria menus in US. The Olive Garden, an Italian family restaurant, has several dishes with meatballs. Meatballs originated in the Chinese Qin dynasty (221 BC to 207 BC). **Question:** Do restaurants associate meatballs with the wrong country of origin? **Rationale:** In what country is the oldest evidence of people eating meatballs found? . . . Are #3 and #1 different? **Answer:** Yes **Facts:** Torah scrolls must be duplicated precisely by a trained scribe. The Torah has a total of 8,674 words. The population of Bunkie Louisiana is 3,939 people according to a 2018 census. **Question:** Can you give at least one word from the Torah to all residents of Bunkie Louisiana? **Rationale:** How many words are in the Torah? How many residents does Bunkie, Louisiana have? Is #1 greater than #2? **Answer:** Yes **Facts:** Wrestlemania X took place in 1994. The Toyota Prius was first manufactured in 1997. **Question:** Could someone have arrived at Wrestlemania X in a Toyota Prius? **Rationale:** When did Wrestlemania X hold? When was the Toyota Prius first manufactured? Is #2 before #1? **Answer:** No

Table 11: Qualitative analysis of exemplars for **StrategyQA** dataset selected by LENS vs EXPLORA. Rationale is not completely shown for some questions to conserve space. However, in our experiments all exemplars include rationales.