Unlocking LLMs' Self-Improvement Capacity with Autonomous Learning for Domain Adaptation

Anonymous ACL submission

Abstract

Self-supervised pre-training and instruction 002 fine-tuning demonstrate the potential of large 003 language models (LLMs) for domain adaptation (DA). In pursuit of superhuman performance, LLMs have demonstrated significant potential in math and coding through selfimprovement algorithms that rely on iterative training with self-generated data. This success stems from the clear reward signals in these environments, which provide a solid foundation for self-improvement. However, when it comes to general DA scenarios, two main challenges 013 emerge: 1) ambiguous self-improvement reward signals and 2) lack of high-quality instruction fine-tuning datasets. This motivates this paper addresses how LLMs can adapt autonomously to new domains using only a large 017 amount of unlabeled target corpora. Inspired by the human practice of self-reflection through open- and closed-book exercises to achieve domain generalization, we propose autonomous 022 learning, which creates a self-improvement learning environment for DA. Here, the model generates questions from documents and conducts two explorations-one with the original document and one with a masked version. By 027 comparing these explorations, the LLMs can independently identify and enhance its policy for reducing knowledge gaps. Experiments across various DA tasks demonstrate that autonomous learning enhances the DA performance of existing models, outperforming traditional finetuning and self-improvement methods.

1 Introduction

034

037

041

Due to the success of self-supervised and instruction tuning methods, Large language models (LLMs) could learn from unsupervised corpora (Kenton and Toutanova, 2019; Qiu et al., 2020; Han et al., 2021), supervised human-annotated instruction data (Ganin and Lempitsky, 2015; Long et al., 2016; Touvron et al., 2023b). Recently, a series of self-improvement methods (Yuan et al., 2024; Chen et al., 2024b) are proposed to enable LLMs to be trained based on its self-generated data, Burns et al. (2023) highlights the challenges of further aligning superhuman models, as their complex behaviors are difficult for humans to effectively supervise. Since the quality of the chain of thought (CoT) can be assessed by the correctness of the final answer (Bai et al., 2022; Wang et al., 2023), a series of self-training methods (Singh et al., 2023; Hosseini et al., 2024; Yang et al., 2024) have been proposed to significantly improve LLMs' performance in math and code. 042

043

044

047

048

053

054

056

060

061

062

063

064

065

066

067

068

069

070

071

072

073

074

076

077

078

079

081

However, when we try to deploy these approaches on general DA scenarios, there are two main challenges that limit the advancement of this field. 1) **Ambiguous self-improvement reward signals**: In general DA problems, the signal used to compare the quality of two responses is implicit. 2) **Lack of high-quality instruction fine-tuning datasets**: The requirement of previous methods for high-quality data further limits the potential of model self-improvement.

It motivates us to study Autonomous Learning in a more practical DA setting, where LLMs adapt to a new domain using only a large amount of target domain unlabeled corpora. In real-world scenarios, humans demonstrate the capacity for Autonomous Learning, such as self-education through reading books or independent research of scientific papers. Most human learning processes are subjective and require minimal guidance, exhibiting strong autonomous characteristics.

To mimic human learning, it reminds us to use Autonomous Learning, an ideal approach to human education. According to (Little, 2002), it is not merely a teaching method; hence, it does not involve teachers dictating behaviors for students to replicate. In (Holec, 1979), the authors define Autonomous Learning as the capacity of learners to direct their own learning, implying their responsibility in shaping various aspects of the learning process. This includes critical thinking, planning, evaluating, and reflecting on learning, with learners actively monitoring the entire process (Benson, 2013). Therefore, autonomous learners are reflective individuals who consciously strive to comprehend what, why, and how they learning (Little, 1996). Consequently, while Autonomous Learning is considered an ideal approach, modern LLM training methods emphasize reliance on human-annotated data and predefined objectives when meet new downstream domain or knowledge, hindering learners' ability to monitor their learning process.

084

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

This inspirs us to adopt **AUTONOMOUS LEARN-ING** for LLMs. The core idea is to enable LLMs to learn autonomously, without human involvement. Autonomous Learning framework provide a selfimprovement environment for DA, therefore, the only prerequisites are the LLMs itself and the learning resources, such as books or documents. The process mimics how a person learns from a book: reading to understand and closing the book to recall and identify areas that require further study to reinforce knowledge. This approach boasts several unique advantages:

- 1. Self-improvement environment in DA. Unlike passive methods, Autonomous Learning involves the model actively engaging with and understanding the material, identifying areas for improvement, and reinforcing its knowledge—emulating the human process of selfimprovement through learning.
- 2. No need for external annotations. As the model undertakes its own learning journey, human intervention becomes unnecessary. The model is fed learning materials such as books, papers, or large corpora—and it dynamically improves itself without the need for annotated data from human, GPT-4 and others.

To assess the efficacy of this learning method, 122 we have set up experiments with learning mate-123 rials of varying scales, such as books (10K para-124 graphs), domain-specific documents (100K para-125 graphs), and Wikipedia (1000K paragraphs), along 126 with corresponding public quizzes to evaluate the 128 learning outcomes. Our experiments demonstrate that Autonomous Learning significantly outper-129 forms pre-training and human-annotated SFT meth-130 ods, suggesting that a model that has diligently 131 'studied' could outperform one that has 'open-book' 132

access but no review. We also introduce recent selfimprovement methods for comparison, and the experiments demonstrate the superiority of our AL's "document in the self-improvement loop." Our findings confirm that Autonomous Learning is a more effective learning method, and its independence from annotations and human involvement significantly reduces the complexity and effort involved in model training.

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

The main contributions of this paper are listed as follows:

- We introduce **Autonomous Learning** for LLMs' DA, a novel training paradigm that introduce a DA self-rewarding environment. This enables LLMs to perform selfimprovement DA without human intervention or other stronger AI, mirroring the natural learning processes of humans.
- We demonstrate that Autonomous Learning eliminates the need for human-annotated data, allowing models to actively engage with and understand learning materials, thereby fostering self-improving learning process.
- Through rigorous experimentation using varied learning materials and corresponding public quizzes, we provide empirical evidence that Autonomous Learning outperforms traditional pre-training, SFT methods, RAG, and self-improvement method.

2 Related Work

In this section, we list some research directions related to this paper. It is important to emphasize that this paper focuses on how to leverage the powerful knowledge and instruction-following capabilities obtained through pre-training and SFT for self-learning within the document to continuously enhance domain adaptability, rather than replacing these techniques. At the end of each part, we will discuss the limitations of each section in the context of further autonomous learning.

2.1 Unsupervised Domain Adaptation

Traditional UDA methodologies encompass174Pseudo-labeling (Ye et al., 2020), the Pivot-based175approach (Pan et al., 2010), and adversarial neural176networks (Ganin et al., 2016). Due to success177of self-supervised learning paradigm's ability178to utilize large-scale unlabeled data, pre-trained179language models (Kenton and Toutanova, 2019;180

278

279

Qiu et al., 2020; Han et al., 2021; Radford et al., 2019) based on self-supervision have become the standard paradigm in unsupervised DA.

181

182

186

187

188

189

190

191

192

193

194

196

197

198

199

203

205

206

Although protocol is concise, such methods face limitations in effectively completing downstream domain adaptation during continuous domain adaptation, because of the lack of differentiated learning strategies for various types of knowledge.

2.2 Supervised Fine-Tuning Domain Adaptation

It has been demonstrated that SFT language models on a collection of datasets expressed in instruction form (Longpre et al., 2023; Touvron et al., 2023b; Yang et al., 2023a) can improve model generalization to unseen tasks, resulting many instructionbased supervised fine-tuning methods (Chung et al., 2024; Touvron et al., 2023a,b) have been introduced. Additionally, a serious of work are proposed to adapt LLMs to structured domain (Ji et al., 2023, 2024) or specific vertical domain such as Chatlaw (Cui et al., 2023), InvestIm (Yang et al., 2023b), Chatharuhi (Li et al., 2023, 2024a).

Although exciting, the SFT method relies heavily on a large amount of high-quality annotations from humans, GPT-4 (OpenAI, 2023), or other sources, posing a formidable barrier to the scalability of instruction tuning practices for larger corpora in the future.

2.3 Self-Training and Self-Improvement

Currently, the methods for self-training and self-211 improvement are mainly developed in the fields of 212 213 math and code. Starting with STaR (Zelikman et al., 2022), reinforced self-training (Gulcehre et al., 214 2023; Zhang et al., 2024), self-rewarding (Yuan 215 et al., 2024; Chen et al., 2024b), focuses on lever-216 aging solutions generated by the LLM to enhance 217 its own performance. These methods involve fine-218 tuning the model on solutions that lead to correct 219 answers. ReST EM (Singh et al., 2023) interprets this fine-tuning as expectation-maximization based 221 reinforcement learning for a solution-generating agent. Discovering successful solutions and how to design the critiquing signal for selecting high quality LLM responses given input queries for further model training are the most challenging problems in self-improvement methods. Early research (Bai et al., 2022; Wang et al., 2023) uses a set of manually created principles or heuristic rules to eliminate low-quality or redundant data. Additionally, Luong 230

et al. (2024) demonstrate that RL-based fine-tuning of an LLM is difficult without initial supervised fine-tuning steps.

The success of these methods is mainly due to the clearly defined reward signals in their selfimprovement loops, which makes them easier to model. In contrast, DA scenarios usually involve numerous unlabeled documents in the target domain, lacking supervisory signals. Even with extensive instruction fine-tuning datasets, the reward signals for self-improvement in general domain adaptation are implicit. Besides, this makes it difficult to apply a unified set of standards to definitively determine whether a knowledge description is True or False.

Unlike previous work, this paper introduces Autonomous Learning (AL) to address the most challenging area of implicit reward signals in the selfimprovement loop for DA. **AL introduces document in the self-improvement loop.** By continuously incorporating external real documents, AL enables the model to access domain-specific knowledge and convert it into trainable data, thus avoiding reliance on self-generated data and preventing model collapse (Shumailov et al., 2024).

3 Preliminary

We define a straightforward learning objective: Given a corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ consisting of *n* documents, and a LLM Φ_{θ} with parameters θ , the goal is to enable Φ_{θ} to effectively learn from this corpus. The effectiveness of this learning can be evaluated using benchmarks related to \mathcal{D} . This process is akin to a person studying a textbook for a course and then being assessed through course exams to gauge their understanding. In our settings, AL only utilizes the source-trained model and unlabeled target data to adapt to the target domain.

4 Methodology: Autonomous Learning

In this section, we provide a detailed implementation of our proposed Autonomous Learning. The overview of our Autonomous Learning framework is shown in Figure 1. This process consists of two stages: **Stage 1. Open-book learning (Warmup):** The model comprehends and absorbs the textual information. **Stage 2. Closed-book learning (Self-Improvement):** The model recalls the content from the first stage, reinforcing and consolidating the learned material. The entire algorithm flow of Autonomous Learning is shown in Algorithm 1.

Stage 1. Open-Book Learning



Figure 1: An ideal learning system should learn autonomously to determine *what to learn, how to learn* and *why to learn*. AL allows for a "document in the self-improvement loop", which allows the model to continuously learn domain corpus autonomously.

Please create a question that closely
aligns with the provided article. Ensure that the <question> does not explicitly reference the text. You may incorporate specific scenarios or contexts in the <question>, allowing the <text> to serve as a comprehensive and precise answer, at the same time, you need to generate an <answer> for the generated <question>. You can refer to the content of the article to answer, but your answer cannot reveal that you have referred to this article. Please output according to the template: '<question>: <answer>:' <document>: [domain-specific document] <question>:</question></document></answer></question></question></answer></text></question></question>

Figure 2: The prompt for document comprehension. [domain-specific document] indicates the document d to be learned.

4.1 Stage 1. Open-Book Learning

283

286

290

Open-book learning simulates the process of studying a book, where we comprehend and absorb its content. The initialization model for Autonomous Learning is a LLM with comprehension abilities, denoted as Φ_{θ^0} . Given a document *d* to be learned, Φ_{θ^0} first comprehends *d* before learning it. This comprehension process can be seen as reading the document and converting it into questions and answers (QA), which can be formalized as: Here, q and a_o represent the questions and answers generated from the document $d_{,,}$ and Prompt refers to the prompt used, as illustrated in Figure 2. For LLMs that cannot follow the prompts, we provide few-shot examples to enable Φ_{θ}^0 to have comprehension abilities, as shown in . In AL, Φ_{θ}^0 first learns from all documents $d \in D$. For documents that are too long, we split them into multiple paragraphs for learning. The objective of open-book learning is:

$$\mathcal{L}_{\text{OpenBook}}(d) = -\log P(a_o | q; \theta^1)$$
 (2)

291

292

294

295

296

297

298

299

300

301

302

303

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

Thus, we obtain the model Φ^1_{θ} after the first stage of learning.

4.2 Stage 2. Closed-book Learning

The model Φ_{θ^1} from the first stage can be thought of as a person who has warm-up a book once. In this process, we usually close the book and recall previously learned content to enhance memory. For the LLM, the second stage involves having the model Φ_{θ^1} recall the learned content without referring to the document, thereby reinforcing the knowledge. We obtain model-generated QA pairs based on *d*:

$$(q, a_o) = \Phi_{\theta^1}(\operatorname{Prompt}(d))$$
 (3)

Note that the questions q generated for the same d vary. For the abstracted questions q from d, Autonomous Learning has the model answer them with the book closed:

$$a_c = \Phi_{\theta^1}(q) \tag{4}$$

$$(q, a_o) = \Phi_{\theta^0}(\operatorname{Prompt}(d)) \tag{1}$$

346 347 348

352

353

354

355

356

357

358

359

360

361

362

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

383

385

386

387

389

390

Input: $\Phi_{\theta^0}, \mathcal{D}$ **Output:** Φ_{θ^2} 1: // Stage 1. Open-Book Learning 2: $\theta^1 \leftarrow \theta^0$ 3: for document d in \mathcal{D} do 4: $(q, a_o) \leftarrow \Phi_{\theta^0}(\text{Prompt}(d))$ // Comprehending document $\begin{array}{l} \ell_1 \leftarrow -\log P(a_o | q; \theta^1) \\ \theta^1 \leftarrow \text{UpdateParameters}(\ell_1, \theta^1) \end{array} \end{array}$ 5: // Absorbing doc-6: ument 7: end for 8: // Stage 2. Close-Book Learning 9: $\theta^2 \leftarrow \theta$ 10: for document d in \mathcal{D} do $(q, a_o) \leftarrow \Phi_{\theta^1}(\operatorname{Prompt}(d))$ 11: $a_c \leftarrow \Phi_{\theta^1}(q) \mathbf{x}$ 12: $\ell_2 \leftarrow -\log \sigma \left(\beta \log \frac{\pi_{\theta^2}(a_o|q)}{\pi_{\theta^1}(a_o|q)}\right)$ $- \beta \log \tfrac{\pi_{\theta^2}(a_c|q)}{\pi_{\theta^1}(a_c|q)} \bigg)$ 13: // Self-reinforcement 14: $\theta^2 \leftarrow \text{UpdateParameters}(\ell_2, \theta^2)$ 15: end for 16: return Φ_{θ^2}

Algorithm 1 The algorithm of Autonomous Learn-

ing

322

324

325

327

329

336

340

where a_c represents the closed-book answers. This gives us a pair (a_o, a_c) . To further explore the online iterative generation of a_c , we conduct experiments in Appendix G. We aim to have the model's closed-book answers $\Phi_{\theta^1}(q)$ approximate a_o as closely as possible. To achieve this, we use a Direct Preference Optimization (DPO) strategy to help the LLM improve the review process. The advantage of DPO is its ability to quickly approximate the correct answers in the presence of biased data. The DPO learning strategy is as follows:

$$\mathcal{L}_{\text{CloseBook}}(d) = -\log \sigma \left(\beta \log \frac{\pi_{\theta^2}(a_o \mid q)}{\pi_{\theta^1}(a_o \mid q)} - \beta \log \frac{\pi_{\theta^2}(a_c \mid q)}{\pi_{\theta^1}(a_c \mid q)} \right)$$
(5)

where $\pi_{\theta^1}(a_c \mid q)$ represents the probability of model Φ_{θ^1} generating a_c given q. In this process, Autonomous Learning treats the open-book answer a_o as the positive answer and the closed-book answer a_c as the negative answer, achieving a selfreinforcing process. See Appendix E for a complete derivation.

5 Experiments

We evaluate our Autonomous Learning (AL) framework across various domains, including commonsense reasoning and domain-specific QA. We compare AL to traditional knowledge injection methods, assess its scalability with different dataset

sizes, and its efficacy in specialized fields like medicine. We also analyze the impact of Open-Book and Closed-Book learning on performance, and evaluate AL's ability of data-efficiency under different low-resource settings.

5.1 Target Domain With Various Scales and Downstream Tasks

To highlight the superiority of our method, we consider the size of the knowledge corpus included in each dataset when selecting them, which varies from 1K to 1M. We train on knowledge corpus and test on multiple downstream tasks corresponding to these specific corpus. The details of our used benchmark is shown in Appendix B.

In all instances, we adopt a prompted zeroshot setup, wherein models are directed to address each task using natural language instructions without any accompanying contextual examples. We choose the more challenging zero-shot setup as we are interested in seeing whether Autonomous Learning works in precisely those cases where a AI system does not specify in advance which instruction should be used in which way for solving a specific problem. In fact, we let the model directly complete downstream tasks to test the model's ability to master knowledge in a specific domain. We use standard greedy decoding. The statistics of these datasets can be found in Table 1. All tasks are measured by accuracy. For tasks under Wiki, we use the reference answers after minor normalization operations mentioned in (Chen et al., 2017; Lee et al., 2019).

5.2 Experiments Setup

Experimental settings. Our research concentrates on unsupervised adaptation scenarios, utilizing Autonomous Learning on an unlabeled target domain corpus to train and enhance an initial model. We hypothesize that a robust model will demonstrate effective generalization and high performance on the target domain's test sets. Our ultimate aim is to transform this model into a domain-specific expert and an instruction model for chat applications, thereby demonstrating the potential of Autonomous Learning in model enhancement and domain-specific adaptation.

Base Model.We use the meta-llama/Llama-2-7b-391chat-hf for experiments, which we call it as initial392model in our experiments.This model originate393

Dataset	Commonsense		Medical				Wiki		
2	OpenBookQA	CNPLE	MedQA-en	MedQA-cn	NQ	TriviaQA	WebQA	TREC	SQuAD
Train	4957	-	10178	27400	78168	78785	3417	1353	78713
Dev	500	-	1272	3425	8757	8837	361	133	8886
Test	500	960	1273	3426	3610	11313	2032	694	10570
	Numbe	r of docu	ments for ea	ich dataset, i	ranging	from 1K t	o 1M		
Documents	1326	87096	156960	163843			1M		

Table 1: The statistical information of the used benchmark.

1394 from HuggingFace ¹.

5.3 Baselines.

396

398

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

427

428

429

430

431

432

To compare with other baselines broadly, we replicate the setups used by prior work and reuse their reported numbers whenever possible. We note that for most tasks, our goal is not to compete with the state-of-the-art (SOTA) because: 1) for tasks like multi-choice and open domain question answering, SOTA models are trained specifically for the corresponding training sets; and 2) SOTA methods often use additional corpora for pretraining that may lead to data contamination, which could confound our domain adaptation studies. We consider the following baselines for our experiments and divide these baselines into two lines: *passive methods* and *autonomous methods*.

For passive methods, we have:

1) **Pre-training**: Following the traditional pretraining paradigms proposed in Kenton and Toutanova (2019); Radford et al.; Tay et al. (2022), we implement a vanilla pre-training method that adopts conventional autoregressive language modeling on given corpora.

2) Supervised Fine-tuning (SFT): We implement 417 a SFT (Ouyang et al., 2022) method named Instruct-418 GPT to perform SFT, which utilizes a substantial 419 amount of manually annotated data, which incurs 420 significant costs. To avoid hallucinations, we use 421 a stronger model to build instructions for a subset 422 of the documents to equip the models with specific 423 instruction following abilities, while we use the 424 tuned model itself to build instructions for the re-425 maining documents. 426

3) Retrieval Augmented Generation (RAG): RAG (Ram et al., 2023) first performs a retrieval step to identify the most relevant document fragments and then fed these documents into the LLMs to serve as the context for generating responses. We retrieve 4 documents for each question. **4) Imbalanced Learning (IL)**: We implement active bias (Chang et al., 2017), a widely used IL method that directly adjust the weights of examples based on the predictive distributions variance. We perform IL on pre-training and supervised fine-tuning, and get 'pre-training + IL' and 'supervised fine-tuning + IL'.

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

For autonomous methods, we have:

1) Self-Tuning: We implement this method (Zhang et al., 2024), in which the model completes data synthesis through self-teaching, and we also use the same amount of the data generated by the stronger model for the warm-up step of instruction following for a fair comparison.

2) SPIN: By automatically generating its own training data and learning from it, SPIN (Chen et al., 2024b) can effectively utilize human-annotated examples for supervised fine-tuning, transforming a weak language model into a powerful one. However, compared to our more rigorous experimental setting, SPIN requires initial annotated data. Therefore, to implement SPIN, we use the self-generated data used for Open-Book learning as the initial real instruction fine-tuning data of SPIN.

For all used LLMs, we use GPT-4 of version gpt-4-0125-preview. Meanwhile, for all methods that require warm-up datas, we construct 1,000 datas using GPT-4 for the commonsense domain and 10,000 datas for others.

5.4 Scaling Laws Across Multi-Magnitude Corpora

As training in deep learning and LLMs becomes increasingly expensive, neural scaling laws can ensure performance. Before training LLMs with hundreds of millions of parameters on massive corpora, we initially train models on smaller-scale corpora and fit scaling laws for training on larger corpora.

Unlike previous work (Henighan et al., 2020; Yang et al., 2023a), which typically fix the size of the corpus and vary the scale of model parameters to observe the effects on error, this paper's scal-

¹https://huggingface.co/

Model	Commonsense		Medcical		Wiki	Avg Acc.
	OBQA	MedQA-cn	MedQA-en	CNPLE	Wiki-5Datasets	
initial model	35.0	26.2	30.5	19.3	38.4	29.9
Passive methods						
Pre-training	37.0	42.6	31.4	30.4	40.2	36.3
Pre-training+IL	38.4	41.8	30.5	27.6	40.2	35.7
RAG	38.4	28.4	26.2	26.0	43.2	32.4
Supervised Fine-Tuning	42.0	52.4	33.2	41.8	42.4	42.4
Supervised Fine-Tuning+IL	41.4	53.3	33.6	42.4	42.5	42.6
Autonomous methods						
Self-Tuning	46.0	54.4	35.1	44.7	43.7	44.8
SPIN	48.4	56.1	36.3	43.1	43.3	45.4
Autonomous Learning (Ours)	53.0	58.2	37.5	46.4	44.6	47.9

Table 2: Results on Common sense, Medical corpora and Wiki corpora. The number of documents has increased from 1,000 to 1,000,000, representing a three-order-of-magnitude growth from the commonsense domain to the Wiki domain. The best performances are highlighted in **bold**, while sub-optimal ones are marked with <u>underline</u>.

ing laws focus more on the corpus. The aim is to demonstrate through experiments on scaling laws of corpora size that our method is universally effective across various scales of corpora. As shown in Table 2, the benchmark results demonstrate that the Autonomous Learning outperforms all the currently most popular knowledge learning paradigms across various document scales. In specific domains such as Medical, the method described in this paper still shows significant improvements.

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489 490

491

492

493

494

Model/Method	MedQA-en	OBQA	CNPLE	
Llama-3.1-8B-Ins	truct			
- initial model	0.386	0.786	0.310	
- SFT	0.405	0.804	0.442	
- SPIN	0.416	0.817	0.456	
- Ours	0.431	0.829	0.481	
Qwen2.5-7B-Instr	ruct			
- initial model	0.335	0.368	0.560	
- SFT	0.366	0.431	0.614	
- SPIN	0.375	0.503	0.631	
- Ours	0.391	0.548	0.678	

Table 3: Experiment of deploying AL on various LLMs as our initial models.

5.5 Effects on Various Models

To highlight the scalability of our method, we deploy our experiments using modern powerful models like Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct in Table 3 Compared to the initial models and other enhancement methods such as SFT and SPIN, our approach consistently achieved the best scores on all testset, demonstrating its ability to enhance model generalization and performance. The consistent performance improvements observed across different models, indicating the strong generalizability of our AL.

5.6 Ablation Study

To better explore the impact of each part of our model, we conducted ablation studies and the results are shown in Table 4. By analyzing the comprehensive ablation experiment settings, we can draw the following conclusions: 1) All ablation models can improve the capabilities of the initial model. 2) Closed-book learning is better than openbook only (ablation model I). 495

496

497

498

499

500

501

502

503

504

505

506

508

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

Furthermore, we find that ablation model IV vield results as expected, even lower than the initial model. One possible explanation is that when removing all terms related to the closed-book answer a_c from the learning objective Formula 5 during the closed-book learning phase, the learning objective of closed-book learning approximates open-book learning. Consequently, training for more epochs leads to overfitting, thereby reducing effectiveness. This finding highlights the effectiveness of AL, wherein self-reflective knowledge contrast further strengthens the model's ability to generalize knowledge. The more detailed experimental results regarding the generalization performance of the Autonomous Learning in two stages are presented in Appendix D. The experimental results indicate that, without the need for additional external annotations, Closed-Book learning can further enhance the domain adaptation performance of existing fine-tuning paradigms.

Interestingly, when we directly perform closedbook learning (the ablation model III), the performance has certain advantages compared to openbook learning, but this effect is still far lower than the complete Autonomous Learning model. The

	Ablation model	OBQA	MedQA-cn	MedQA-en	CNPLE
-	initial model	35.0	26.2	30.5	19.3
Ι	open-book only	40.0	51.4	32.4	40.5
Π	closed-book only	44.4	52.6	33.7	42.3
III	closed-book \rightarrow open-book	48.4	54.3	35.2	44.1
IV	AL w/o a_c in closed-book	33.6	25.4	28.3	19.6
VI	open-book \rightarrow closed-book (AL)	53.0	58.2	37.5	46.4

Table 4: Ablation study. Ablation model III represents training first with the closed-book method, followed by the open-book method.

	OBQA	MedQA-cn	MedQA-en	CNPLE
initial model	35.0	26.4	30.5	19.3
SFT	42.0	50.3	33.0	40.8
AL				
- full Doc.	53.0	58.2	37.5	46.4
- fewer Doc.				
# 30%	50.2	56.9	36.6	45.6
# 15%	44.2	52.4	35.3	43.3
# 5%	38.6	51.6	34.2	39.5

Table 5: Low-resource settings where it adopts fewer documents in Autonomous Learning (AL).

reason may be due to the lack of learning of all documents by the model in the open-book learning stage. As a result, when closed-book learning is performed directly, although the model's learning method based on self-knowledge comparison can learn a certain amount of knowledge, it is still under-fitting.

To demonstrate that AL is not dependent on warm-up data, we use few-shot prompting to enable the llama-2 model to generate D->QA instruction fine-tuning data independently. We then conducted experiments based on the model's self-synthesized warm-up data. Table 6 show that AL can consistently output all baseline models.

Model/Method	MedQA-en	OBQA	CNPLE
Llama-2-7b-chat			
 initial model 	30.5	0.350	19.3
- SFT	31.4.	0.420	41.8
- SPIN	36.3	0.484	43.1
- AL	37.5	0.530	46.4
- AL w/o warm-up	<u>36.7</u>	<u>0.514</u>	<u>44.7</u>

Table 6: Performance without warm-up dataset. We still provide the necessary warm-up data to baselines.

5.7 Competitive Performance Achieved by Fewer Documents

The Closed-Book phase of our approach aims to enhance the model's generalization of learned knowledge and can be seamlessly integrated into any model that has undergone the Open-Book learning phase to further enhance its learning effectiveness. To investigate the knowledge enhancement effects of our approach in the Closed-Book learning phase, we conducted an in-depth exploration of the relationship between model performance and the quantity of documents used for reinforced knowledge learning in this phase.

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

Table 5 illustrates the experimental results of our approach in the Closed-Book phase under different scales of document subsets. It can be observed that our approach in the Closed-Book phase demonstrates performance comparable to the full dataset when based on only 30% of the documents. Additionally, when only 5% of the documents are available, our approach rapidly enhances the model's generalization of knowledge, achieving performance on par with SFT.

This highlights the efficient utilization of documents by our approach, which can extract rich knowledge through self-learning even with a small number of documents, thereby enhancing the model's generalization of knowledge.

6 Conclusion

In this paper, we explore the significant challenges associated with enabling LLMs to autonomously adapt to new domains by leveraging extensive unlabeled target corpora. We propose and validate **Autonomous Learning**, which innovatively introduces a self-improvement environment for DA. By enabling LLMs to self-educate through direct interaction with diverse textual materials, this approach not only mimics human learning processes but also significantly enhances the capabilities of LLMs beyond the constraints of traditional training methods reliant on human-annotated data. With the help of sufficient pre-training and SFT, our results show that AL outperforms all baselines without any additional human annotations.

530

531

532

533

534

535

685

686

687

688

635

Limitations

588

591

592

594

595

596

597

605

611

612

613

615

616

617

618

619

625

629

630

631

634

Despite its promising performance in three domain adaptation tasks, AL has several limitations that must be considered:

> • Limited Autonomous Learning Data Format: AL focuses on the most practically significant domain adaptation setting, where the target domain has a large amount of unlabeled data. It explores how to use these datas for domain adaptation through a self-improvement paradigm. However, current AL methods only focus on text modality and its unlabeled corpora. In future work, AL should support more diverse multimodal domain adaptation scenarios.

• Additional Computational Cost: Although AL can further push the boundaries of domain adaptation beyond existing methods, it requires two inferences per step in closed-book learning, which increases the overall training time. This suffers from the same shortcomings as recent self-training-based methods, such as ReST (Gulcehre et al., 2023),selfrewarding (Yuan et al., 2024), self-play (Chen et al., 2024b). In future research, simpler AL methods need to be explored to improve the training efficiency of the AL framework.

• Limited to Models with Instruction-Following Capabilities: The method of this paper starts directly from an initial model, which needs to have sufficient instructionfollowing capabilities to complete both openbook and closed-book answers. However, for models that do not possess this instructionfollowing capability like GPT-2 (Radford et al., 2019), we can use chat models like Llama-2-7b-chat-hf (Touvron et al., 2023b), Baichuan 2-Chat-7b (Yang et al., 2023a), ChatGPT (OpenAI, 2022) to simply construct instruction fine-tuning datasets to enable them to master the instruction-following required for Autonomous Learning.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

- Petr Baudiš and Jan Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 222–228. Springer.
- Phil Benson. 2013. *Teaching and researching: Autonomy in language learning*. Routledge.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.
- Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer opendomain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879.
- Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024a. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.
- Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.
- Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

796

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domainadversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

693

700

701

704

705

706

707

710

711

712

715

716

717

718

720

721

722

723

724

725

727

728

729

733

734

735

736

737 738

739

740

741

742

- Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced selftraining (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.
- Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.
- Henri Holec. 1979. Autonomy and foreign language learning. ERIC.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-star: Training verifiers for self-taught reasoners. arXiv preprint arXiv:2402.06457.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023.
 C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint* arXiv:2305.08322.
- Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2918– 2933.
- Ke Ji, Peng Wang, Wenjun Ke, Guozheng Li, Jiajun Liu, Jingsheng Gao, and Ziyu Shang. 2024. Domainhierarchy adaptation via chain of iterative reasoning for few-shot hierarchical text classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6315–6323. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. pages 1601–1611.

- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. pages 6086–6096.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.
- David Little. 1996. The politics of learner autonomy. *Learning Learning*, 2(4):7–10.
- David Little. 2002. Autonomy in language learning: Some theoretical and practical considerations. In *Teaching modern languages*, pages 89–95. Routledge.
- Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning.
- Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.
- OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.
- OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

801

- 807 810 811 812 813 814 815 816 817 818
- 819

820

821

- 822 826 827
- 830 831 832
- 836

838

- 839 840
- 841 842
- 847

- 852

- Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the 19th international conference on World wide web, pages 751-760.
- Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10):1872-1897.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.
 - Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.
 - Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In Thirty-seventh Conference on Neural Information Processing Systems.
 - Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. pages 2383-2392.
 - Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11:1316–1331.
 - Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. Nature, 631(8022):755-759.
- Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2023. Beyond human data: Scaling self-training for problem-solving with language models. arXiv preprint arXiv:2312.06585.
- Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. 2022. Ul2: Unifying language learning paradigms. arXiv preprint arXiv:2205.05131.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

853

854

855

856

857

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

889

890

891

892

894

895

896

897

898

899

900

901

902

903

904

905

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13484-13508.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122.
- Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. arXiv preprint arXiv:2309.13064.
- Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7386-7399
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. arXiv preprint arXiv:2401.10020.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. Neural Information Processing Systems (NeurIPS).
- Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. 2024. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching. arXiv preprint arXiv:2406.06326.

909

910

911

912

A Ethics Statement

B

for test data.

The datasets used in this study are all derived from publicly available resources on the internet and are freely accessible. And the backbone models we use are also publicly available. Therefore, there is no need for ethics concern.

and Downstream Tasks

sponding downstream tasks.

Target Domain With Various Scales

Below we describe each domains and its corre-

Commonsense: We choose a small-scale corpus

dataset in the domain of common sense, Open-

BookQA, which contains a corpus of 1,326 com-

mon sense entries to serve as reference knowledge

• OpenBookQA (OBQA) (Mihaylov et al.,

2018) comprises 5,957 multiple-choice ques-

tions, each offering four possible answers.

The dataset is combined with external funda-

mental scientific facts. To successfully answer

these questions, one must have a comprehen-

sive understanding of these fundamental sci-

Medical: We pick three widely used datasets in

Medical domain. Each dataset is accompanied by

a medical textbook, which contains the knowledge

required to answer the questions in the dataset. We

split the textbook corpus into multiple documents,

each containing no more than 512 tokens. After

dividing the textbooks, the CNPLE, MedQA-en,

and MedQA-cn datasets contain 87,096, 156,960,

and 163,843 documents, respectively. Please note

that MedQA-cn and CNPLE are written in Chinese.

• MedQA-en (Jin et al., 2021) gathers ques-

tions from the National Medical Board Ex-

aminations of the USA. MedQA presents a

entific facts. and its applications.

- 913 914
- .
- 915
- 916 917
- 918 919
- 920
- 92
- 923
- 924
- 926
- 927
- 928 929

92

930

- 931 932
- 933
- 935 936

937

938 939

ç

941

- 0
- 945 046
- 947

948

949

950

951

952

953

- demanding benchmark because it incorporates diverse medical knowledge—including patient profiles, disease symptoms, and drug dosage requirements. This variety requires contextual understanding for accurately answering the questions posed.
- MedQA-cn (Jin et al., 2021) is also collected from the National Medical Board Examinations of the Mainland China. For both MedQA-en and MedQA-cn, we test them on the 4-option questions.

• The 2023 Chinese National Pharmacist Licensure Examination (CNPLE) (Chen et al., 2023) is a fresh medical exams. Addressing data contamination in the training of Large Language Models (LLMs) is challenging, particularly when dealing with complex and vast datasets (Huang et al., 2023). To mitigate this issue, we use the 2023 Chinese National Pharmacist Licensure Examination, conducted on October 21, 2023, as our benchmark. The release date of this dataset is later than all the base and chat models we used, therefore it can prevent data leakage and ensure reliable evaluations.

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

Wiki: We use the same five QA datasets and training/dev/testing splitting method as in previous work (Lee et al., 2019). For datasets under this part, we train on the documents in Wiki corpus as their common corpus. Here, we select a subset of the Wikipedia corpus that contains 1 million documents.

- Natural Questions (NQ) (Kwiatkowski et al., 2019) was designed for end-to-end question answering. The questions were mined from real Google search queries and the answers were spans in Wikipedia articles identified by annotators.
- **TriviaQA** (Joshi et al., 2017) contains a set of trivia questions with answers that were originally scraped from the Web.
- WebQuestions (WQ) (Berant et al., 2013) consists of questions selected using Google Suggest API, where the answers are entities in Freebase.
- CuratedTREC (TREC) (Baudiš and Šedivý, 2015) sources questions from TREC QA tracks as well as various Web sources and is intended for open-domain QA from unstructured corpora.
- **SQuAD v1.1** (Rajpurkar et al., 2016) is a popular benchmark dataset for reading comprehension. Annotators were presented with a Wikipedia paragraph, and asked to write questions that could be answered from the given text.

We collectively refer to these datasets as Wiki-5Datasets in out experiments.

	Hyperparameters	OpenBookQA	CNPLE	MedQA-en	MedQA-en	wiki
	Optimizer			AdamW		
	Warmup Ratio			0.1		
	Learning Rate			2e-5		
Open-Book Stage	LR Schedule			cosine		
	Batch Size			8		
	Max Length			2048		
	# Epoch			3		
	Optimizer			Rmsprop		
	Warmup Ratio			0.2		
	Learning Rate			5e-7		
	LR Schedule			Linear		
Closed-Book Stage	Batch Size			8		
	Max Length			2048		
	DPO beta			0.01		
	# Epoch			3		

Table 7: The hyperparameters used for Our Autonomous Learning on all benchmark.

C Hyperparameters of Autonomous Learning

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

The training hyperparameters of Autonomous Learning on different datasets are reported in Table 7. For all of the hyperparameters, we directly use the same value across all datasets. The training was conducted on a GPU server with 8 NVIDIA A100 GPU cards.

D Naive Empirical Risk Minimization is Not Enough

In this section, we emphasize the point of this paper, that Naive Naive Empirical Risk Minimization (EMR) is not enough, through trend charts on various datasets. In Figures 3, it can be observed that all Naive EMR methods exhibit clear plateaus, and additional epoch training does not yield higher performance but rather leads to overfitting. The closed-book learning method introduced in the second stage of this paper further enhances the model's domain adaptation, resulting in improved accuracy for the corresponding tasks, indicating the effectiveness of the knowledge-contrasting approach proposed in this paper.

E Mathematical Derivations of AL

In this appendix, we will clarify that our approach is a process of autonomously enhancing domain adaptation based on knowledge comparison, rather than simply praising or criticizing. We propose the advantages of RL methods in two ways.

First, by (Rafailov et al., 2023) Section 4, the

gradient of DPO loss is:

$$\nabla_{\theta} \mathcal{L}_{DPO} = -\beta E_{(x, y_w, y_l) \sim D} [\sigma(\hat{r}_{\theta}(x, y_l) - 103)]$$

1031

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

(6)

$$\hat{r}_{\theta}(x, y_w))(\nabla_{\theta}\log \pi_{\theta}(y_w|x) - \nabla_{\theta}\log \pi_{\theta}(y_l|x))]$$
 10

where (x, y_w) and (x, y_l) are the chosen and rejected responses, respectively. The updated parameters of the model will move in the direction making the difference $\nabla_{\theta} \log \pi_{\theta}(y_w|x) - \nabla_{\theta} \log \pi_{\theta}(y_l|x)$ become larger with a weight function $\sigma(\hat{r}_{\theta}(x, y_l) - \hat{r}_{\theta}(x, y_w))$, not just increase the log probability of the chosen one and decrease the log probability of the rejected one. Actually in (Rafailov et al., 2023), it has been shown that if we just increase the chosen probability and decrease the rejected probability, the language model will degenerate. Our experiment (Figure 5) shows that the rewards of chosen and rejected responses can be increase or decrease simultaneously.

Second, by Equation (4) in (Rafailov et al., 2023), the optimal solution of the KL-constrained reward maximization objective is:

$$\pi(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x) \exp(\frac{1}{\beta}r(x,y)) \quad (7)$$

for the given reference model π_{ref} and reward r, 1052 where Z(x) is the normalization factor independent 1053 of the responses. Hence we can see that the optimal 1054 solution is not just choose the best response and ig-1055 nore all other ones, it is distributed to all responses 1056 with the probability determined by the reward func-1057 tion and β , higher reward leads to higher probabil-1058 ity. It can be seen that for two different responses 1059 y_1, y_2 , although there is a better one, but if they are 1060 both good enough, that means $r(x, y_1)$ and $r(x, y_2)$ 1061



Figure 3: The performance gap between open-book learning and closed-book learning. Epoch 0 stands for the performance of initial model. Epochs 4 to 6 represent the 1st, 2nd, and 3rd epochs of closed-book learning, respectively.

are closed with each other, there probabilities in the optimal distribution will be closed. So the RL methods for the LLM training is not just praising or criticizing, but only depends on their actually rewards. Responses with high reward values will have high probabilities in the end.

1062

1064

1065

1066

1067

1069

1071

1072

1073

1075

1079 1080

1081

1082

1084

F Demonstrating How Autonomous Learning Works Through Examples

In this appendix, we demonstrate how Autonomous Learning works through some examples. As shown in Figure 6, we observe that after one epoch of closed-book learning, the closed-Book answer in Epoch 2 aligns better with the learned documents and questions that the closed-book answer in Epoch 1.

G Impact of Online Iterative Data Generation

To explore the online iterative generation of a_c , we designed an experiment where, after training the model with k samples, we regenerate a_c and a_o for subsequent training based on the updated model. We tested various values of k in (4096, 16,384, 65,536, and 131,072) to observe the performance

trends. The experiment was conducted using the llama-2-7b-chat-hf model.

We observe that when the update frequency is high (i.e. when k is 4096), the model's performance actually deteriorates. Conversely, the model performs best when k is set to 65,536. However, increasing k to 131,072 does not lead to further improvements. The possible reason for this is that a high update frequency implies the model uses relatively less data for training in each iteration, which may cause it to over-fit the most recently observed data, thus affecting its generalization ability and leading to unstable learning. As we gradually re-

Model/Method	MedQA_en	CNPLE
initial model	0.305	0.193
offline AL (Ours)	0.375	0.464
online AL		
- k=4096	0.352	0.446
- k=16384	0.369	0.463
- k=65536	0.383	0.479
- k=131072	0.364	0.456

Table 8: The impact of online iterative data generation

1095

1096

1097

1085





(b) Logits for the closed-book answer.

Figure 4: The trend of logits variation for open-book answers and closed-book answers on the MedQA-cn dataset.



Figure 5: Reward performance on MedQA-cn dataset.

duce the update frequency, we find a performance inflection point that surpasses offline AL. This observation suggests that finding the optimal balance between data efficiency and learning stability is crucial. Future research should focus on designing a more effective online AL strategy to achieve this balance.

<|system|>

1098

1099

1100

1101

1102

1103

1104

You are KnowledgeGPT, equipped with in-depth knowledge. Your task is to directly answer the user's question. <|user|>

[question] What is an example of a film that combines comedy with a critique of human vices through interwoven stories, highlighting the impact of greed among others? <|assistant|>

[answer] An example of a film that skillfully combines comedy with a critique of human vices, specifically focusing on the theme of greed among others...

Table 9: An example of our QA instruction tuning data.

Epoch 1	
<pre>{Reference document} An example of hunting is an otter cracking open clams with a rock {Question} What is an example of how otters use tools in their dietary habits? {Open-book answer} An example of how otters use tools in their dietary habits is by using rocks to crack open clams. This behavior demonstrates their ability to adapt and manipulate their environment to access food, showcasing their intelligence and problem-solving skills. {Closed-book answer} Otters are known to use tools in their dietary habits by using rocks to crack open clams and other shellfish. They have been observed holding the shells in their paws and then slamming them on rocks to break</pre>	Please create a question the vided article. Ensure that plicitly reference the text. scenarios or contexts in a <text>to serve as a comprete the same time, you need to generated <question>. Yee the article to answer, but yee you have referred to this at to the template ' <question>: <answer>:'<document>:[domain-secific document the domain-specific document the domain-specific pre-treated</document></answer></question></question></text>
them open. This behavior demonstrates their problem-solving skills and adaptability in their dietary habits.	
them open. This behavior demonstrates their problem-solving skills and adaptability in their dietary habits.	
<pre>them open. This behavior demonstrates their problem-solving skills and adaptability in their dietary habits.</pre>	

Figure 6: An example to show how Autonomous Learning works. Please create a question that closely aligns with the provided article. Ensure that the <question>does not explicitly reference the text. You may incorporate specific scenarios or contexts in the <question>, allowing the <text>to serve as a comprehensive and precise answer, at the same time, you need to generate an <answer>for the generated <question>. You can refer to the content of the article to answer, but your answer cannot reveal that you have referred to this article. Please output according to the template ' <question>:... <answer>:....' <document>: [domain-specific document] <question>:: <answer>:

Figure 7: The prompt for question generation. [domain-specific document] refers to a document in the domain-specific pre-training corpora.

You are KnowledgeGPT, equipped with in-depth knowledge. Your task is to directly answer the user's question. <question>: [question generated by LLM] <answer>:

Figure 8: The prompt for the answer generation of QA. [question generated by LLM]] is the previously textderived query in Figure 7.