

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

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are widely applied in real world scenarios, but fine-tuning them comes with significant computational and storage costs. Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA mitigate these costs, but the adapted parameters are dependent on the base model and cannot be transferred across different backbones. One way to address this issue is through knowledge distillation, but its effectiveness inherently depends on training data. Recent work such as TransLoRA avoids this by generating synthetic data, but this adds complexity because it requires training an additional discriminator model. In this paper, we propose **TiTOK**, a new framework that enables effective LoRA Transplantation through **Token**-level knowledge transfer. Specifically, TiTOK captures task-relevant information through a **token-wise** contrastive excess between a source model with and without LoRA. This excess highlights informative tokens and enables selective filtering of synthetic data, all without additional models or overhead. Through experiments on three benchmarks across multiple transfer settings, our experiments show that **TiTOK** is consistently effective, achieving average performance gains of + 4–8% compared to baselines overall.

1 INTRODUCTION

Large Language Models (LLMs) (Brown et al., 2020; Vaswani et al., 2023), have made significant progress in many real-world applications, including chatbots (OpenAI et al., 2024), search engines (Xiong et al., 2024), and coding assistants (Rozière et al., 2024). While fine-tuning LLMs has been demonstrated to be a promising way to improve performance on downstream tasks, it incurs substantial computational and storage costs. Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019) methods such as LoRA (Hu et al., 2021) alleviate this burden by updating only a small subset of parameters while keeping the base model frozen. However, PEFT’s adapted parameters are dependent to the base model and cannot be transferred across different models. This limitation is increasingly critical given the rapid release of new LLMs and the growing diversity of available models.

One considerable approach to mitigate this limitation is through Knowledge distillation (KD) (Hinton et al., 2015; Azimi et al., 2024), which transfers the knowledge embedded in a source model’s PEFT adapters to a target model with new PEFT adapters by aligning the target’s output distributions with those of the source. However, KD is inherently data-dependent, typically requiring access to training data from target downstream tasks (Nayak et al., 2019; Liu et al., 2024), which is often unavailable or costly to obtain. To address this limitation, TransLoRA (Wang et al., 2024) has recently proposed using synthetic data by leveraging the data synthesis capabilities of recent LLMs (Wang et al., 2023; Kim et al., 2025). This approach enables the target model to acquire domain knowledge without direct access to the original dataset. Nevertheless, TransLoRA requires training an additional discriminator model to filter low-quality synthetic data, which inevitably introduces extra complexity and computational overhead. Furthermore, it primarily emphasizes the role of synthetic data, paying less attention to *how the knowledge transfer process itself should be designed*.

Contribution. In this paper, we propose a new framework that enables effective LoRA Transplantation through **Token**-level knowledge transfer (**TiTOK**). Our high-level idea is to selectively convey task-relevant information from the source model’s LoRA by using token-level signals to guide the transfer process, rather than relying on the entire token sequence. We specifically capture this information through a concept we introduce as **token-wise contrastive excess**, obtained by

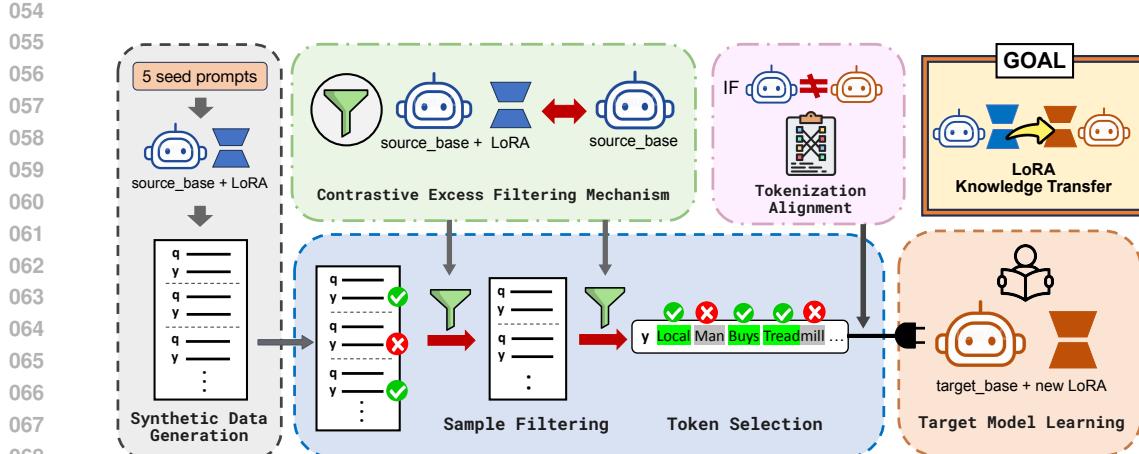


Figure 1: Overview of **TiTOK**: Transplantation through Token-level knowledge transfer. Starting from a small set of seed prompts, the source expert model (source base model + LoRA) generates synthetic data. A **token-wise** contrastive excess filtering mechanism then compares the expert against its base backbone to compute token-level excess scores. Using these scores, TiTOK first performs sample filtering and subsequently token selection, retaining only the most informative samples and tokens. When tokenizers differ, masks are aligned prior to training. The resulting filtered data is finally used to train a new LoRA on the target backbone, enabling efficient knowledge transfer.

comparing predictions from a source model with LoRA and the same model without LoRA. Intuitively, this **token-wise** contrastive excess highlights tokens that contain important task knowledge. This excess signal is further utilized to filter the **generated** synthetic data for training, thereby enabling selective learning on samples that contain richer information. Unlike TransLoRA, which requires **training** an additional discriminator model for filtering, TiTOK requires neither extra models nor additional training overhead. Moreover, we design an effective mechanism to resolve tokenizer mismatches between source and target models, which enhances robustness and applicability **TiTOK**.

We demonstrate the effectiveness of TiTOK by conducting extensive experiments on three widely used benchmarks, which cover both reasoning (Big Bench Hard (Suzgun et al., 2022) and MMLU (Hendrycks et al., 2021)) and personalization tasks (LaMP (Salemi et al., 2024)). In particular, TiTOK improves the performance by +7.96% over the vanilla target model, +6.0% over KD, and +4.4% over TransLoRA, when averaged across all tasks and transfer settings. We also explored a variety of transfer settings, including transfers within the same model, across different model families and sizes, and even across different model versions. In every case, our approach achieves consistent improvements. Interestingly, TiTOK remains effective even when applied to external data originating from tasks different from the target task, highlighting both robustness and general applicability. Overall, these empirical results highlight TiTOK as a methodologically simple yet powerful paradigm for efficiently transferring LoRA knowledge across models in diverse scenarios.

2 RELATED WORKS

Transferring PEFT adapters. Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019; Li & Liang, 2021) has emerged as both a practical and popular alternative to full model fine-tuning. By requiring updates to only a small portion of parameters, it enables efficient adaptation, with LoRA (Low-Rank Adaptation) (Hu et al., 2021; Dettmers et al., 2023) standing out as one of the most widely adopted methods. However, a fundamental limitation is that LoRA adapters are tied to the frozen backbone they were trained on, making them difficult to transfer to other base models. To address this issue, recent studies such as TransLoRA (Wang et al., 2024) have attempted to transplant the knowledge of LoRA adapters across models by generating synthetic data (Wang et al., 2023). While effective to some extent, this approach requires an additional discriminator model to filter high-quality synthetic data, resulting in a relatively heavy pipeline. In parallel, Knowledge

108 Distillation (KD) (Hinton et al., 2015; Azimi et al., 2024) has been widely explored as another
 109 means of knowledge transfer; however, traditional KD typically operates within a teacher–student
 110 framework at the logit or sequence level and requires access to training data close to the original,
 111 in order to use the teacher’s distribution for supervising the student. In contrast, our approach
 112 focuses on **token-level selective transfer**, offering a more fine-grained yet lightweight alternative
 113 that enables efficient transplantation of LoRA adapters across diverse models for deployment.
 114

115 **Data synthesis with LLMs.** Synthetic data generation with LLMs has attracted increasing attention
 116 as a means to reduce reliance on costly or inaccessible datasets (Wang et al., 2023; Kim et al.,
 117 2025). Prior lines of research have leveraged synthetic data for purposes such as privacy preservation
 118 (Bu et al., 2025), data augmentation (Kumar et al., 2021), and domain adaptation (Li et al.,
 119 2023). In our framework, synthetic data serves as a core component, enabling the transfer of LoRA
 120 adapters without access to the original training corpus, while simultaneously mitigating privacy
 121 concerns and reducing the degree of dependency on external datasets. Moreover, since both queries and
 122 labels are generated directly by the source expert model itself, synthetic data ultimately provides a
 123 self-sufficient mechanism that fits to our objective of lightweight and effective knowledge transfer.
 124

125 **Selective token training.** Recent studies (Lin et al., 2025; Gu et al., 2020) demonstrate that not
 126 all tokens contribute equally to model training, motivating research on selective training strategies.
 127 While such approaches have primarily been applied to accelerate optimization or reduce redundancy
 128 (Yeongbin et al., 2024; Bal et al., 2025), our work innovatively extends this concept to the setting
 129 of knowledge transfer. Specifically, our concept of excess scores (Eq. 2) is derived from this idea,
 130 where the contrast between the source backbone and its LoRA adapter yields token-level judgments.
 131 This enables TiTOK to transplant LoRA knowledge in a more focused and fine-grained manner,
 132 highlighting the broader applicability of selective training beyond its original scope.
 133

3 TiTOK: TRANSPLANTING LoRA THROUGH TOKEN-LEVEL KNOWLEDGE

135 In this section, we introduce TiTOK, a framework for LoRA Transplantation through **Token**-level
 136 knowledge transfer (Fig. 1). The core idea of TiTOK is to transfer the knowledge from a source
 137 model’s LoRA adapter into a target model’s LoRA adapter by training specifically on the infor-
 138 mative tokens within synthetic data. Specifically, the framework consists of three components: (1)
 139 *Synthetic data generation* (Sec. 3.1), where the source expert model produces query–label pairs for
 140 target task; (2) *Excess score computation* (Sec. 3.2), which calculates token-level importance using
 141 source model; and (3) *Target model training with filtering* (Sec. 3.3), which trains target model with
 142 newly initialized LoRA adapter on top-ranked samples and tokens. In addition, we propose *Excess
 143 score alignment* (Sec. 3.4), an algorithm designed to apply TiTOK even when tokenizers of the
 144 source and target models differ. The overall algorithm of TiTOK is presented in Alg. 1.
 145

3.1 SYNTHETIC DATA GENERATION VIA LLM PROMPTING WITH FEW-SHOT DATA

147 Let \mathcal{M}_s denote the source backbone LLM and \mathcal{A}_s its LoRA adapter on target downstream task,
 148 which forms the source expert model $\mathcal{M}_s + \mathcal{A}_s$. The target model, whose LoRA adapter \mathcal{A}_t will be
 149 trained, is denoted by \mathcal{M}_t . Then, TiTOK first constructs a synthetic dataset \mathcal{D}_s , similar to the idea of
 150 TransLoRA (Wang et al., 2024). This usage of synthetic data allows us to avoid keeping the whole
 151 original dataset for the downstream task, and simultaneously let \mathcal{A}_t learn the knowledge encoded
 152 in the source adapter \mathcal{A}_s . Unlike TransLoRA’s approach of using the untuned target model \mathcal{M}_t to
 153 generate synthetic data, we use the source expert model $\mathcal{M}_s + \mathcal{A}_s$ to synthesize data (see empirical
 154 comparison in Fig. 2). Concretely, $\mathcal{M}_s + \mathcal{A}_s$ synthesizes both the *query* and the *label* within a
 155 prompting-based data synthesis framework (Wang et al., 2023) (see details in Appendix R); given a
 156 few-shot data of downstream task, it first generates a query \mathbf{q} , and then produces the corresponding
 157 label \mathbf{y} conditioned on \mathbf{q} . To encourage diversity, we apply ROUGE-L filtering together with dedu-
 158 plication to all tasks, except for exceptional cases where such filtering is infeasible (more details are
 159 in Appendix L). Consequently, the resulting synthetic dataset consists of query–label pairs:
 160

$$\mathcal{D}_s = \{(\mathbf{q}_j, \mathbf{y}_j)\}_{j=1}^N. \quad (1)$$

161 In this way, \mathcal{M}_t would be trained on the synthetic data \mathcal{D}_s generated by the source expert model
 $\mathcal{M}_s + \mathcal{A}_s$, thereby enabling knowledge transfer without relying on the entire original dataset.

162 3.2 TOKEN-WISE CONTRASTIVE EXCESS SCORE FROM SOURCE MODEL WITH LORA
163

164 As described in Sec. 3.1, TiTOK relies on synthetic data, but synthetic data is often prone to
165 imperfections (Chen et al., 2024b); thus, sophisticated filtering is essential to retain high-quality
166 informative samples. TransLoRA (Wang et al., 2024) tackles this challenge with a separate discrimi-
167 nator to filter useful queries, but this introduces the extra burden of training and maintaining an extra
168 model. In contrast, we propose a lightweight alternative that only utilizes the already trained source
169 model. Specifically, we use source model and its LoRA adapter to perform two complementary
170 roles: (1) the *amateur* role (\mathcal{M}_s) and (2) the *expert* role ($\mathcal{M}_s + \mathcal{A}_s$). Then, the difference between
171 the two roles provides an *implicit supervision signal* where the task information is encoded.

172 Formally, let $\mathbf{y} = [y_1, \dots, y_L]$ denote the synthesized response, where y_i is a token of \mathbf{y} corresponding
173 to the synthesized query \mathbf{q} . Then, we define the *excess score* as:

$$174 \quad S(y_i) = L_e(y_i) - L_a(y_i), \quad (2)$$

175 where the amateur and expert losses on token y_i are defined as:

$$177 \quad L_a(y_i) = \log P_{\mathcal{M}_s}(y_i \mid \mathbf{q}, \mathbf{y} < i), \quad L_e(y_i) = \log P_{\mathcal{M}_s + \mathcal{A}_s}(y_i \mid \mathbf{q}, \mathbf{y} < i). \quad (3)$$

178 The excess score $S(y_i)$ quantifies the knowledge discrepancy incurred by equipment of LoRA,
179 thereby identifying the tokens where the adapter provides a decisive contribution. Intuitively, if
180 the backbone model is uncertain about predicting a token but the LoRA-enhanced model assigns it
181 with high confidence, that token will thus obtain a large excess score. This implies that tokens with
182 higher $S(y_i)$ correspond to positions where the LoRA adapter injects task-specific knowledge that
183 the backbone could not capture on its own. In this way, the excess score functions as a *fine-grained*
184 *attribution signal*, derived entirely from the internal behavior of the model itself, and guides training
185 toward the specific regions of data that are most enriched with the adapter’s knowledge.

187 3.3 TARGET MODEL TRAINING WITH SAMPLE FILTERING AND TOKEN SELECTION
188

189 After the computation of excess scores $S(y_i)$, the newly initialized LoRA adapter \mathcal{A}_t for target
190 model \mathcal{M}_t is trained using the synthetic samples $(\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_f$ with two-level filtering schemes.

191 **First stage: Sample filtering.** We begin by filtering the synthetic dataset \mathcal{D}_s (Eq. 1) at the sample
192 level to remove less informative examples. For each synthetic sample, we compute the mean of the
193 excess scores $S(y_i)$ across the tokens in \mathbf{y} and retain only M samples with the highest values.

$$194 \quad \bar{S}_j = \frac{1}{|\mathbf{y}_j|} \sum_{y_i \in \mathbf{y}_j} S(y_i). \quad (4)$$

197 Let \mathcal{D}_f be the set of the M samples in \mathcal{D}_s with the largest \bar{S}_j :

$$199 \quad \mathcal{D}_f = \text{Top}M\{ (\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_s : \bar{S}_j \}. \quad (5)$$

200 Through this step, the synthetic data undergoes a filtering process, ensuring that subsequent training
201 is concentrated specifically on the remaining examples in \mathcal{D}_f with richer knowledge signals.

202 **Second stage: Token selection.** Next, we consider token selection; that is, \mathcal{A}_t does not learn from
203 all tokens within the retained samples. Instead, it focuses only on those prioritized by the excess
204 scores $S(y_i)$, which are identified as most important for knowledge transfer. To achieve this, we
205 select the top $k\%$ of tokens ranked by their excess scores using the indicator $I_{k\%}(y_i)$:

$$207 \quad I_{k\%}(y_i) = \begin{cases} 1, & \text{if } \text{rank}_{\mathbf{y}_j}(S(y_i)) \leq \lfloor k\% \cdot |\mathbf{y}_j| \rfloor, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

209 where $|\mathbf{y}_j|$ denotes the number of tokens in \mathbf{y}_j , and $\text{rank}_{\mathbf{y}_j}(S(y_i))$ indicates the rank of $S(y_i)$ among
210 the tokens of that response. Based on this selection, the training objective for \mathcal{A}_t is defined as

$$212 \quad \mathcal{L}_{\text{TiTok}} = \sum_{(\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_f} \sum_{y_i \in \mathbf{y}_j} I_{k\%}(y_i) \cdot L_t(y_i), \quad (7)$$

214 where $L_t(y_i)$ is the negative log-likelihood loss assigned by $\mathcal{M}_t + \mathcal{A}_t$ on token y_i (only \mathcal{A}_t is
215 learnable). By training only on these filtered tokens (Eq. 7), TiTOK enables \mathcal{A}_t to efficiently acquire
the source LoRA’s knowledge without access to the original training data or any external models.

216 **Algorithm 1:** TiTOK: Transplanting LoRA through Token-Level Knowledge

217 **Input:** source expert $\mathcal{M}_s + \mathcal{A}_s$, target \mathcal{M}_t , parameters $N, M, k\%$

218 **Output:** trained target LoRA \mathcal{A}_t

219

220 1. Construct synthetic dataset $\mathcal{D}_s = \{(\mathbf{q}_j, \mathbf{y}_j)\}_{j=1}^N$ with $\mathcal{M}_s + \mathcal{A}_s$

221 2. **for** $(\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_s$ **do**

222 | Compute token excess scores $S(y_i) = L_e(y_i) - L_a(y_i)$

223 | Calculate mean score $\bar{S}_j = \frac{1}{|\mathbf{y}_j|} \sum_{y_i \in \mathbf{y}_j} S(y_i)$

224 3. Select top- M samples by \bar{S}_j to form \mathcal{D}_f

225 4. **for** $(\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_f$ **do**

226 | Rank tokens by $S(y_i)$ and keep top- $k\%$, represented by mask $I_{k\%}(y_i)$

227 5. **if** $\text{tokenizer}(\mathcal{M}_s) \neq \text{tokenizer}(\mathcal{M}_t)$ **then**

228 | Align masks $I_{k\%}^{(s)}(y_i) \rightarrow I_{k\%}^{(t)}(y_i)$

229 6. Train \mathcal{A}_t on \mathcal{M}_t with masked loss $\mathcal{L}_{\text{TiTok}} = \sum_{(\mathbf{q}_j, \mathbf{y}_j) \in \mathcal{D}_f} \sum_{y_i \in \mathbf{y}_j} I_{k\%}(y_i) \cdot L_t(y_i)$

230

231 **return** \mathcal{A}_t

232

233

234 3.4 EXCESS SCORE ALIGNMENT ACROSS DIFFERENT TOKENIZERS

235 In cases of transfer between models with different tokenizers, a direct mapping of token-level
 236 signals is not possible, given that the source and target models may segment text differently. To
 237 address this, we introduce a simple yet robust tokenizer alignment algorithm that propagates token
 238 masks (Eq. 6) from the source token sequence $\mathbf{y}^{(s)}$ to the target token sequence $\mathbf{y}^{(t)}$. The algorithm
 239 first aligns token sequences using dual pointers that incrementally decode and match text spans.
 240 Masks are then propagated using the following four rules: (1) direct copy for one-to-one mappings,
 241 (2) replication for one-to-many, (3) averaging for many-to-one, and (4) averaging with replication
 242 for many-to-many. Finally, a top- $k\%$ selection step retains the most confident target tokens. This
 243 process ensures consistent supervision across tokenizers, enabling reliable transfer even when
 244 models tokenize text differently. The conceptual illustration of this procedure is presented in Fig.4.

245 4 EXPERIMENT

246 In this section, we present our experimental results to answer the following research questions:

- 247 • **RQ1:** Can TiTOK efficiently transfer knowledge of LoRA in various scenarios? (Table 1)
- 248 • **RQ2:** What is the contribution of each component in TiTOK? (Table 2)
- 249 • **RQ3:** How sensitive is token-level selective transfer to the selection ratio? (Figure 3)
- 250 • **RQ4:** How does the choice of model to synthesize query affect the performance? (Figure 2)
- 251 • **RQ5:** Can TiTOK transfer knowledge using data from a different or unrelated domain? (Table 3)

252 4.1 EXPERIMENTAL SETUPS

253 **Models.** We mainly present our knowledge transfer experiments using models from the Mistral
 254 and Llama families. Specifically, we designed the following LoRA transfer (*source* \rightarrow *target*) setups.
 255 (1) Mistral-7B-Inst-v0.3¹ \rightarrow Mistral-7B-Inst-v0.3: the basic transfer setup, (2) Mistral-7B-Inst-v0.3
 256 \rightarrow Llama-3.1-8B-Inst²: the different-family model transfer setup, (3) Llama-3.2-3B-Inst³ \rightarrow
 257 Llama-3.1-8B-Inst: the different-size model transfer setup, and (4) Llama-2-7b-chat-hf⁴ \rightarrow Llama-
 258 3.1-8B-Inst: the different-version model transfer setup. These setups are intended to test whether a
 259 smaller model can effectively transfer knowledge to a larger one, and to explore whether a relatively
 260 weaker model can still influence a newer, stronger model. These various setups realistically reflect
 261 how LLMs develop today, where various models are released and newer, improved versions keep
 262

263 ¹<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

264 ²<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

265 ³<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

266 ⁴<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

270 emerging. Model names are abbreviated for clarity and are used consistently in all tables and
 271 figures: 1) “Mistral 7B” = Mistral-7B-Instruct-v0.3, 2) “Llama3 8B” = Llama-3.1-8B-Instruct, 3)
 272 “Llama3 3B” = Llama-3.2-3B-Instruct, 4) “Llama2 7B” = Llama-2-7b-chat-hf.
 273

274 **Baselines.** To demonstrate the effectiveness of TiTOK, we compare it against three baselines: (i)
 275 *Vanilla*, the target base model without any fine-tuning; (ii) *KD(+MinED)*, knowledge distillation
 276 (KD) from source expert model (Hinton et al., 2015). When the source and target models use dif-
 277 ferent tokenizers, the original KD is not applicable. In this case, we use Minimum-Edit-Distance
 278 (MinED) tokenizer alignment (Wan et al., 2024), which aligns both token sequences and vocabulary
 279 distributions via dynamic-programming sequence alignment for near matches (e.g., “gets” \leftrightarrow “get”)
 280 and by mapping probability mass to nearest edit-distance neighbors (e.g., “immediately” \leftrightarrow “imme-
 281 diate”). We use the synthesized data by TransLoRA as the training data for KD and MinED ; and
 282 (iii) *TransLoRA* (Wang et al., 2024), a prior method in which the vanilla target model synthesizes the
 283 queries, while the source model with its LoRA adapter generates the corresponding synthetic labels.
 284 A discriminator is then employed to filter the synthetic data for training the target LoRA adapter.
 285

286 **Datasets.** Following prior work in TransLoRA, we first conduct experiments on two representative
 287 benchmarks: (1) *Big-Bench Hard (BBH)* (Suzgun et al., 2022) and (2) *Massive Multitask Language*
 288 *Understanding (MMLU)* (Hendrycks et al., 2021). BBH consists of 27 challenging reasoning tasks
 289 structured as multiple choice or short answer questions, designed to test compositional generaliza-
 290 tion and advanced problem solving abilities of a Language Model. Meanwhile, MMLU covers 57
 291 tasks across diverse academic subjects, presented in multiple choice format to evaluate broad knowl-
 292 edge and reasoning skills of a model. Since both benchmarks only provide test sets, we split the data
 293 into 90% for training the source expert model $\mathcal{M}_s + \mathcal{A}_s$ and the remaining 10% for evaluation.
 294

295 To extend our approach to personalization and text generation, we additionally conduct experiments
 296 on LaMP benchmark (Salemi et al., 2024), focusing exclusively on its generation tasks. In particular,
 297 we experiment with (3) *News Headline Generation (LaMP 4)* and (4) *Scholarly Title Generation*
 298 (*LaMP 5*), as they are the only text generation tasks that are both accessible and reliably evaluable.
 299 The remaining LaMP tasks are excluded, given that LaMP 1-3 are discriminative, and LaMP 6-7
 300 lack gold labels. For LaMP tasks, the source expert model $\mathcal{M}_s + \mathcal{A}_s$ is trained on data from the 30
 301 users with the longest activity histories. From each user, we use 200 data points for training and 50
 302 data points for validation, a design choice intended to conduct a more rigorous and robust evaluation.
 303 In total, each LaMP task contains 6,000 training examples and 1,500 evaluation examples.
 304

305 To assess performance on the BBH and MMLU, we measure the average accuracy using the
 306 LM-Eval Harness (Gao et al., 2024). Following the setup used in TransLoRA (Wang et al., 2024),
 307 all tasks are conducted in a zero-shot setting. Meanwhile, for the LaMP tasks, we adopt ROUGE-1
 308 and ROUGE-L scores as evaluation metrics, following the benchmark’s primary evaluation metrics.
 309

310 **Implementation details.** For training of both source and target models, we use a learning rate
 311 of 5×10^{-5} , train for 2 epochs with a batch size of 4, and apply LoRA with rank $r = 8$, scaling
 312 factor $\alpha = 8$, and dropout 0.05. Optimization is performed using AdamW with weight decay
 313 1×10^{-2} , together with a linear learning rate schedule and a warmup ratio of 0.1. For synthetic
 314 data generation, we provide five samples from the original training data as few-shot exemplars,
 315 and apply top-p sampling (Holtzman et al., 2020) to generate both queries and labels, with sampling
 316 hyperparameters tuned individually for each task. To further encourage diversity, we apply ROUGE-
 317 L filtering with a threshold of 0.7 and deduplication to remove redundant queries, following Wang
 318 et al. (2023). For tasks where ROUGE-based filtering is infeasible, we apply only deduplication (see
 319 Appendix L). For the initial synthetic pool, we generate $2M$ synthetic samples and, after filtering
 320 with token-wise contrastive excess scores, retain the top M , where M equals the source training set
 321 size (see Appendix K). For token selection, the selection ratio $k\%$ is fixed at 70% across all tasks
 322 and transfer settings, except for the Llama3 3B \rightarrow Llama3 8B transfer setting, where we find that
 323 $k\% = 30\%$ consistently yields the best performance. **When inferencing during the evaluation stage,
 324 we adopt greedy decoding to ensure fully deterministic and reproducible inference results.**

325 4.2 MAIN RESULTS

326 Table 1 summarizes the experimental results on BBH, MMLU, and LaMP (News Headline and
 327 Scholarly Title generation) across four transfer settings. First, we observe that transfer within

324 **Table 1: Main results.** Experiments on BBH, MMLU, News Headline and Scholarly Title Generation
 325 tasks under four transfer settings. BBH and MMLU are reasoning tasks and are evaluated using
 326 LM-Eval Harness, while News Headline and Scholarly Title Generation represent personalization
 327 tasks and are evaluated with ROUGE-1/L. All evaluations are zero-shot. Best scores are in **bold**.
 328

Transfer	Method	BBH	MMLU	News Headline		Scholarly Title	
		Acc.	Acc.	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
Mistral 7B → Mistral 7B	Vanilla	0.397	0.557	0.117	0.101	0.381	0.311
	KD	0.426	0.556	0.121	0.107	0.392	0.320
	TransLoRA	0.424	0.533	0.155	0.136	0.447	0.382
	TiTOK (ours)	0.432	0.563	0.160	0.142	0.473	0.414
Mistral 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD+MinED	0.477	0.484	0.127	0.112	0.455	0.389
	TransLoRA	0.471	0.472	0.122	0.108	0.461	0.397
	TiTOK (ours)	0.482	0.488	0.140	0.124	0.464	0.403
Llama3 3B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD	0.470	0.475	0.126	0.111	0.449	0.383
	TransLoRA	0.460	0.466	0.121	0.107	0.455	0.387
	TiTOK (ours)	0.509	0.475	0.127	0.113	0.457	0.392
Llama2 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378
	KD+MinED	0.473	0.478	0.125	0.111	0.450	0.384
	TransLoRA	0.472	0.468	0.123	0.109	0.453	0.388
	TiTOK (ours)	0.510	0.479	0.140	0.122	0.461	0.404

348 the same model family is highly effective; when transplanting the LoRA adapter trained on the
 349 Mistral 7B into a fresh instance of the same model, TiTOK consistently surpasses all baselines. In
 350 particular, TiTOK improves the vanilla model by 24.08% on average across all tasks, and further
 351 outperforms the KD and TransLoRA baselines by 19.63% and 4.91%, respectively. Taken together,
 352 the findings show that, as a first step, transfer within the same family is reliably successful.

353 Beyond same family transfer, we also find that TiTOK is highly effective in cross-model transfer
 354 settings. For example, when transferring from Mistral 7B to Llama 8B, TiTOK delivers an average
 355 improvement across all tasks of 7.11% over the vanilla model, while outperforming KD by 4.73%
 356 and TransLoRA by 6.24%. This highlights that TiTOK is not confined to intra-family knowledge
 357 transfer, but can successfully bridge across architectures. In the case of Llama3 3B to Llama3
 358 8B, TiTOK yields average gains of 3.45% against vanilla, 2.49% against KD, and 4.14% against
 359 TransLoRA. These results suggest that TiTOK scales effectively with model size, making it useful
 360 when moving from lightweight to larger models without losing efficiency. Finally, when transferring
 361 from Llama2 7B to Llama3 8B, TiTOK performs average advantages of 7.43% over vanilla, 6.28%
 362 over KD, and 7.22% over TransLoRA. This indicates that TiTOK adapts robustly even across different
 363 model versions, implying practical relevance when models are upgraded in real-world pipelines.

364 Collectively, these results provide strong evidence that TiTOK not only excels in transfers within
 365 the same model family, but also demonstrates broad effectiveness in cross-model transfers, thereby
 366 highlighting its robustness across a wide spectrum of model scales, versions, and families.

367 4.3 ADDITIONAL ANALYSES

369 **Ablation study.** To validate the contribution of each component in our framework, we conduct ad-
 370 dditional experiments by selectively excluding the sample filtering and token selection mechanisms in
 371 Sec. 3.3. We report the average performance scores across all four transfer settings, and the results
 372 are presented in Table 2. When sample filtering is not applied (1st and 2nd rows), the training data
 373 for the target model is randomly sampled from the full synthetic dataset. Under this setting, apply-
 374 ing only token selection (2nd row) performs noticeably better than the purely random baseline. This
 375 demonstrates the effectiveness of our token selection approach and empirically confirms that **token-
 376 wise** contrastive excess reliably identifies and selects the most informative tokens. Similarly, when
 377 only sample filtering is applied (1st and 3rd rows), the results improve over pure random sampling,
 indicating that selecting high-quality data is essential. This finding further underscores that incorpo-

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
Table 2: **Ablation study.** "Sample filtering" uses the mean `token-wise` contrastive excess score to remove uninformative examples, while "Token selection" further refines the retained data by selectively keeping only the most informative tokens within each sample. Without sample filtering, data are randomly sampled. Results are reported as accuracy (Acc.) for BBH and MMLU, averaged across tasks. Meanwhile, we report ROUGE-1 (R-1) and ROUGE-L (R-L) for LaMP tasks.

Settings		BBH	MMLU	News Headline		Scholarly Title	
Sample filtering	Token selection	Acc.	Acc.	R-1	R-L	R-1	R-L
✗	✗	0.458	0.485	0.133	0.117	0.456	0.393
✗	✓	0.463	0.496	0.137	0.121	0.460	0.397
✓	✗	0.470	0.500	0.139	0.122	0.460	0.397
✓	✓	0.483	0.501	0.142	0.125	0.464	0.403

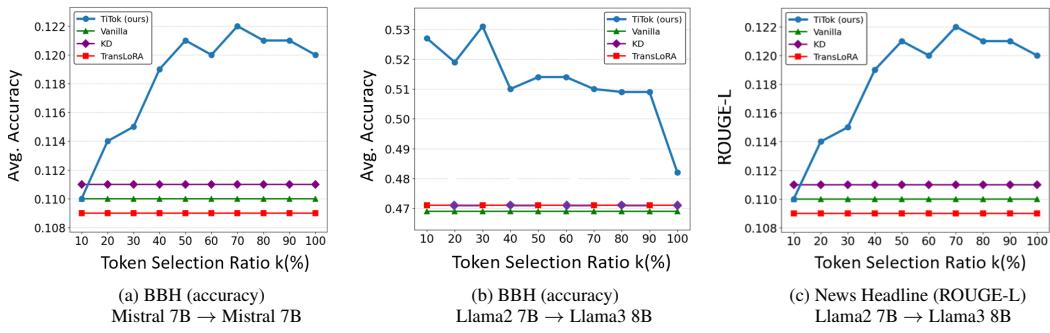


Figure 3: **Representative performance trends across $k\%$.** Among the three graphs, two come from the same task (BBH), while two share the same transfer setting (Llama2 7B → Llama3 8B).

rating an effective filtering mechanism for synthetic data is essential. In our framework, `token-wise` contrastive excess serves this role by reliably selecting samples with richer and more informative signals, as demonstrated clearly by the empirical results. Finally, the results show that combining both stages (4th row) achieves the best performance, confirming their complementary roles in filtering high-quality examples and selecting informative tokens for effective knowledge transfer.

Impact of query generation model. We now move on to examine how the choice of query generation model influences performance. Practically, synthetic queries can be generated either by the source model or by the target model, with the latter being the approach originally adopted in TransLoRA's pipeline (Wang et al., 2024). Figure 2 compares these two options across different transfer settings, with the reported scores averaged over all BBH tasks. Notably, we observe that using the source expert model (source backbone + LoRA) for query generation generally yields stronger performance than using the target model. One possible explanation is that when both the synthetic queries and the corresponding labels are generated by the same model, they remain closer to its training distribution. This alignment between queries and labels likely makes the supervision more coherent and accurate, thereby facilitating more effective transfer. These results suggest that maintaining distributional alignment between queries and labels is a key factor for improving knowledge transfer, thereby providing empirical justification for our design choice of using the source model as both the synthetic query and label generator.

Effect of token selection ratio. We now proceed to explore the impact of token selection ratios ($k\%$) in our framework. Fig. 3 presents three representative curves: one task (BBH) under two

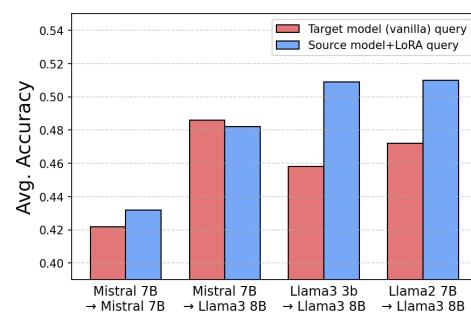


Figure 2: **Impact of query source.** Across transfer settings, using the source expert model to synthesize query generally yields better performance. The accuracy averaged over all BBH tasks are reported.

transfer settings, and another task (News Headline Generation) under the same transfer setting for comparison. For BBH with the same backbone on source and target (Mistral 7B→Mistral 7B), a moderate token selection ratio of 40-70% yields the best results. Very low ratios (10-20%) lead to underfitting, while a 100% ratio is ineffective as it applies no filtering at all. Interestingly, in a weak-to-strong transfer (Llama2 7B → Llama3 8B) for the same BBH task, performance improves as $k\%$ decreases. This suggests that selecting only the tokens with the highest excess loss effectively filters out the majority of noisy regions where the weaker model is uncertain. By discarding the majority of tokens where the weaker model lacks confidence, TiTOK can significantly reduce negative transfer. This trend, however, reverses for the News Headline Generation task under the same weak-to-strong transfer setting. For this task, a larger $k\%$ is generally better, potentially because even a weaker model can provide valuable lexical and stylistic cues that are useful for personalization. Consequently, unlike reasoning-focused tasks such as BBH, which are sensitive to noisy supervision, personalization tasks like News Headline Generation benefit more broadly from the source model’s outputs, even when the source model is relatively weaker than the target model.

Effectiveness of transfer through external data source. We further examine whether TiTOK can transfer knowledge effectively in cases where synthetic data is not preferred, and thus external data is used as an alternative. To this end, we evaluate three alternative settings on LaMP tasks for Mistral 7B→Mistral 7B transfer setup: (1) using data from a randomly chosen different user, (2) mixing data from multiple users, and (3) transferring across tasks, where data of Scholarly Title Generation is used to train on News Headline Generation and vice versa (*i.e.*, out-of-distribution scenario). The results are presented in Table 3. Remarkably, TiTOK consistently outperforms all baselines across these heterogeneous external settings. These findings demonstrate that TiTOK is not restricted to synthetic data scenarios, but can also adapt effectively under external or user-provided data conditions. This highlights the flexibility of TiTOK for practical deployment in diverse real-world application scenarios and further underscores its adaptability across different data conditions, as it is effective even when external data is used as an alternative to synthetic data.

5 CONCLUSION

In this paper, we propose TiTOK, an efficient framework that transfers LoRA knowledge from a source model to a target model by training only on a selectively chosen set of highly informative tokens. At its core, TiTOK leverages token-level signals to distill task-relevant information from the source adapter, rather than relying on the entire token sequence. By focusing supervision on the regions where the adapter contributes most, TiTOK achieves stronger and more targeted knowledge transfer. With this simple yet effective design, TiTOK proves robust across tasks and consistently surpasses existing baselines, making it a practical solution for efficient knowledge transfer.

Limitations and Future Directions While TiTOK has shown robustness and clear advantages over existing methods, there remain opportunities to refine and extend the framework. Firstly, although TiTOK is designed to minimize dependence on original data, it still requires a small number of seed examples to generate synthetic data via prompting. Nevertheless, this reliance is modest, as we avoid using full datasets. In addition, our experiments confirm that external data can also serve as an effective alternative, reinforcing the flexibility of the approach. Regarding token selection, TiTOK currently applies a fixed threshold to determine which tokens are retained. While this simple design is effective and stable across tasks, future work could explore more adaptive or data-driven thresholding strategies to further enhance efficiency without compromising robustness.

486 ETHICS STATEMENT
487488 We conduct this research in full accordance with established ethical standards. For our experiments
489 we rely exclusively on publicly available datasets such as LaMP, MMLU, and BBH and use them
490 strictly in accordance with their intended purpose for academic research. For the LaMP tasks, which
491 involve user data, our TITOK framework aligns with ethical considerations by minimizing data
492 dependence. It does not store or expose raw user data and only updates a small set of task and
493 user-specific parameters. This helps safeguard privacy while enabling efficient knowledge transfer.
494495 REPRODUCIBILITY STATEMENT
496497 We provide a comprehensive description of our implementation in Section 4, including pipeline
498 configurations, hyperparameters, models, datasets, and evaluation metrics. The source code for our
499 implementation and experiments will be made publicly available in a repository upon publication.
500501 REFERENCES
502503 Rambod Azimi, Rishav Rishav, Marek Teichmann, and Samira Ebrahimi Kahou. Kd-lora: A
504 hybrid approach to efficient fine-tuning with lora and knowledge distillation. *arXiv preprint*
505 *arXiv:2410.20777*, 2024.507 Melis Ilayda Bal, Volkan Cevher, and Michael Muehlebach. Eslm: Risk-averse selective language
508 modeling for efficient pretraining, 2025. URL <https://arxiv.org/abs/2505.19893>.510 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari-
511 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal,
512 Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.
513 Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz
514 Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec
515 Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL
516 <https://arxiv.org/abs/2005.14165>.517 Hyungjune Bu, Chanjoo Jung, Minjae Kang, and Jaehyung Kim. Personalized llm decoding via
518 contrasting personal preference, 2025. URL <https://arxiv.org/abs/2506.12109>.519 Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay
520 Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with
521 fewer data, 2024a. URL <https://arxiv.org/abs/2307.08701>.523 Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay
524 Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data.
525 In *International Conference on Learning Representations (ICLR)*, 2024b.527 Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning
528 of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.530 Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Fos-
531 ter, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muen-
532 nighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang
533 Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model
534 evaluation harness, 07 2024. URL <https://zenodo.org/records/12608602>.535 Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. Train no evil: Se-
536 lective masking for task-guided pre-training. In Bonnie Webber, Trevor Cohn, Yulan He, and
537 Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language
538 Processing (EMNLP)*, pp. 6966–6974, Online, November 2020. Association for Computational
539 Linguistics. doi: 10.18653/v1/2020.emnlp-main.566. URL <https://aclanthology.org/2020.emnlp-main.566/>.

540 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Ja-
 541 cob Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.

542

543 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*
 544 *preprint arXiv:1503.02531*, 2015.

545

546 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text
 547 degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

548

549 Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, An-
 550 drea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp.
 551 In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

552

553 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
 554 and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.

555

556 Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril
 557 Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. Evaluating language models
 558 as synthetic data generators. In *Annual Meeting of the Association for Computational Linguistics
 (ACL)*, 2025.

559

560 Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained trans-
 561 former models, 2021. URL <https://arxiv.org/abs/2003.02245>.

562

563 J. Li and S. Babu. *A Graduate Course on Statistical Inference*. Springer, 2019.

564

565 Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv*
 566 *preprint arXiv:2101.00190*, 2021.

567

568 Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large lan-
 569 guage models for text classification: Potential and limitations, 2023. URL <https://arxiv.org/abs/2310.07849>.

570

571 Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu
 572 Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need, 2025.
 573 URL <https://arxiv.org/abs/2404.07965>.

574

575 He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge
 576 distillation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

577

578 Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and
 579 Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *Proceedings of the
 580 International Conference on Machine Learning (ICML)*, 2019.

581

582 Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hy-
 583 potheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337,
 584 1933.

585

586 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
 587 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red
 588 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-
 589 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher
 590 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-
 591 man, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann,
 592 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis,
 593 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey
 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,
 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila
 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,
 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-
 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan

594 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-
 595 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan
 596 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,
 597 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun
 598 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-
 599 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook
 600 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel
 601 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen
 602 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel
 603 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,
 604 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv
 605 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,
 606 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,
 607 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel
 608 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-
 609 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,
 610 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel
 611 Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe
 612 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,
 613 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,
 614 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra
 615 Rimbach, Carl Ross, Bob Rotstetd, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,
 616 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-
 617 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,
 618 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,
 619 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,
 620 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-
 621 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-
 622 jayvergyia, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan
 623 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,
 624 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-
 625 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming
 626 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao
 627 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL
 628 <https://arxiv.org/abs/2303.08774>.

629 Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi
 630 Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémie Rapin, Artyom Kozhevnikov, Ivan Ev-
 631 timov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong,
 632 Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier,
 633 Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.
 634 URL <https://arxiv.org/abs/2308.12950>.

635 Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large lan-
 636 guage models meet personalization, 2024. URL <https://arxiv.org/abs/2304.11406>.

637 Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung,
 638 Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-
 639 bench tasks and whether chain-of-thought can solve them, 2022. URL <https://arxiv.org/abs/2210.09261>.

640 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,
 641 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.

642 Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion
 643 of large language models, 2024. URL <https://arxiv.org/abs/2401.10491>.

644 Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini, Aude Oliva, Rogerio Feris, and
 645 Leonid Karlinsky. Trans-lora: towards data-free transferable parameter efficient finetuning. In
 646 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

648 Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and
649 Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions.
650 In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

651

652 Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and
653 Sumi Helal. When search engine services meet large language models: Visions and challenges,
654 2024. URL <https://arxiv.org/abs/2407.00128>.

655 Seo Yeongbin, Dongha Lee, and Jinyoung Yeo. Train-attention: Meta-learning where to focus in
656 continual knowledge learning. *Advances in Neural Information Processing Systems*, 37:58284–
657 58308, 2024.

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

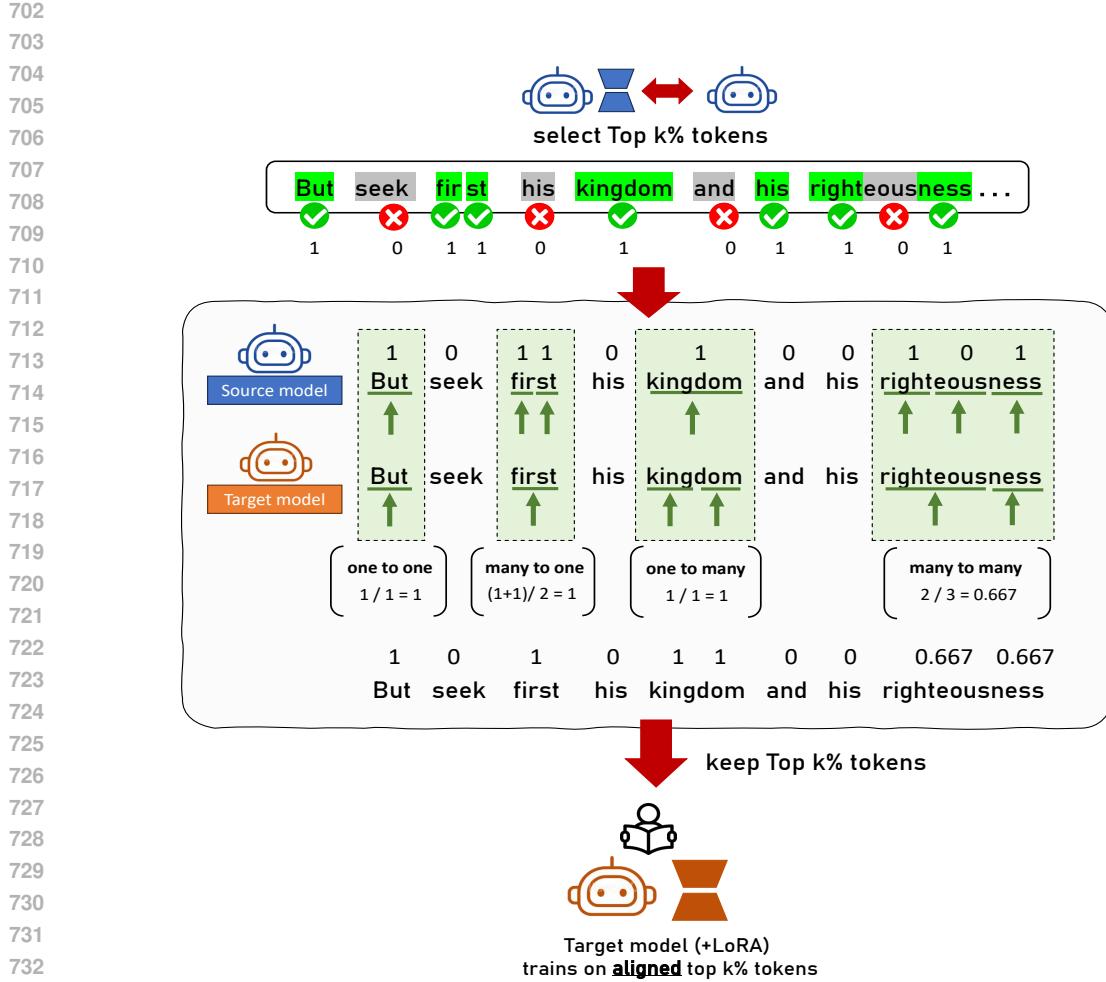


Figure 4: Overview of **TiTok’s tokenizer alignment algorithm**. The algorithm handles cases where the source and target models use different tokenizers. The binary mask scores assigned by the source model are averaged within aligned spans and propagated to target tokens, producing fractional scores that guide top- $k\%$ token selection for training the target model’s LoRA adapter.

A TOKENIZER ALIGNMENT ALGORITHM

When the source and target models use different tokenizers, direct token-level transfer is not possible. To address this, we implement a dual-pointer alignment procedure. The algorithm maintains two pointers, one for the source tokens and one for the target tokens. At each step, the source pointer advances by one token, accumulating a decoded segment, while the target pointer incrementally extends its own segment until the normalized texts match. Once a match is found, the corresponding spans are recorded as an alignment, and the target pointer jumps forward to that position. After alignment, we apply masking rules to propagate the source binary mask scores (values that indicate whether a token should be kept or discarded) to the target tokens. Specifically:

- 1) **One-to-One**: binary mask score is directly copied.
- 2) **One-to-Many**: score is replicated across all aligned target tokens.
- 3) **Many-to-One**: averaged scores of multiple source tokens are assigned to the target token.
- 4) **Many-to-Many**: the averaged score of the source tokens is assigned to the corresponding aligned target tokens.

756
757 Table 4: Average alignment percentage breakdown across BBH and News Headline tasks in the
758 Mistral 7B → Llama3 8B transfer setting.
759

	BBH	News Headline
exceptions	0%	0%
many to many	46.18%	5.00%
many to one	3.47%	17.45%
one to many	0.01%	0.01%
one to one	50.33%	77.52%

760
761
762
763
764
765 Table 5: Performance comparison for Mistral-7B → Llama-3 8B transfer.
766
767
768

		BBH	News Headline	
		Acc.	R-1	R-L
Mistral-7B → Llama-3 8B	Vanilla	0.397	0.117	0.101
	One to one only	0.472	0.138	0.120
	TiTOK (k=70%)	0.482	0.160	0.142

776 This process yields fractional mask scores that capture the relative importance of each target token.
777 Finally, we keep only the top- $k\%$ of tokens according to these scores, producing a final binary
778 selection mask over the target tokens that preserves the most informative regions while discarding
779 less relevant ones. An overview of this tokenizer alignment algorithm is presented in Figure 4.
780

781 **Robustness of the algorithm.** The algorithm is conceptually error-free since it simply performs
782 a deterministic mapping between two tokenizers on the exact same text sequence. Since both to-
783 kenizations correspond to the identical underlying character string, the mapping cannot introduce
784 semantic errors; every token in both tokenizers is defined over non-overlapping spans of the same
785 text, and these spans align uniquely. For instance, in the Mistral 7B → Llama3 8B transfer setting
786 on BBH and News Headline benchmarks, 100% of tokens are aligned, and there are no exception
787 cases as shown in Table 4, confirming that alignment errors do not occur in practice.
788

789 Furthermore, we additionally construct a degraded setting in which we keep only the one-to-one
790 aligned token pairs and discard all other alignment cases. We perform the experiment in the Mistral
791 7B → Llama3 8B transfer setting on BBH and News Headline tasks, and the full results are presented
792 in Table 5. From these empirical results, we can observe that using all the full alignment cases
793 outperforms using one-to-one token pairs only. This implies that the many-to-one, one-to-many,
794 and many-to-many alignments are also correctly aligned, as including them yields the best results.
795 Together, these results confirm that our alignment method is both accurate and robust.
796

797 B THEORETICAL FOUNDATIONS AND EMPIRICAL VALIDATION OF 798 TOKEN-WISE CONTRASTIVE EXCESS

800 In this section, we present the theoretical and empirical evidence supporting both our token-wise
801 contrastive excess scoring and our knowledge transfer mechanism. The proposed token-wise
802 contrastive excess score is not a heuristic but a token-level log-likelihood ratio (LLR) between the
803 source expert (backbone + LoRA) and the backbone. LLR is a robust metric widely established
804 in statistical testing (Li & Babu, 2019) for example, by the Neyman–Pearson lemma (Neyman &
805 Pearson, 1933), the likelihood ratio is known to be the optimal statistic for identifying differences
806 between two models. Therefore, high-LLR tokens are exactly the regions where the adapter changes
807 the predictive distribution, i.e., where task knowledge is injected. Consequently, selecting high-LLR
808 tokens is theoretically justified as extracting the most informative adapter-specific signals.
809

In addition, from the perspective of standard knowledge distillation, tokens with near-zero LLR
provide no additional teacher knowledge and offer no benefit for transfer. In contrast, high-LLR

Transfer	Method	Subgroup	Acc.
Mistral 7B → Llama3 8B	Vanilla	–	0.469
		top 20%	0.482
	TiTok	bottom 20%	0.468
		random 20%	0.476

Table 6: Effect of token subgroup selection in Mistral 7B → Llama3 8B transfer setting. Selecting the **top 20%** tokens by token-wise contrastive excess scoring yields the best performance in BBH.

tokens correspond to maximal teacher–student divergence, larger gradients, and higher information contribution to the student. Therefore, emphasizing tokens with high LLR in particular naturally focuses training on the precise positions where transferable knowledge is concentrated.

We verify this empirically by contrasting token subgroups of the top 20%, bottom 20%, and random 20% on the BBH in the Mistral 7B → Llama3 8B transfer setting. To prevent unintentional overlap across groups and to extract the most informative region, we employ a comparatively low 20% threshold for this analysis. The results are presented in Table 6. Analytically, the results show that the top 20% produces the best transfer performance, thereby clearly validating that tokens with high token-wise contrastive excess scores indeed contain concentrated task knowledge.

Furthermore, we demonstrate that truly important tokens are consistently selected regardless of the expert’s model. For example, for BBH datasets, Mistral 7B and Llama2 7B experts agree on 59.76% of the chosen tokens even when we limit the selection to a small top 20% subset to prevent trivial overlap. This implies that a token will continue to be recognized across models if it is truly significant for the task, and it can be effectively identified by our token-wise contrastive excess score.

C DETAILS OF DATASETS

C.1 BIG-BENCH HARD (BBH)

Big-Bench Hard (BBH) (Suzgun et al., 2022) is designed as a rigorous benchmark for evaluating model performance on challenging reasoning problems, including multi-step logical reasoning, symbolic manipulation, and commonsense inference. The tasks are formatted as multiple choice questions. Since BBH is originally a test-only benchmark, we split 90% of the data for training the source expert model and reserved 10% for evaluation. The 27 BBH tasks are categorized in Table 7.

C.2 MASSIVE MULTITASK LANGUAGE UNDERSTANDING (MMLU)

Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) is a comprehensive benchmark for evaluating model performance across a broad range of knowledge intensive tasks. The benchmark consists of multiple choice questions and, similarly to BBH, we also apply a 90%/10% split of the original test-only data. All the 57 subtasks are categorized in Table 8.

C.3 LAMP TASKS

We utilize the LaMP benchmark to evaluate whether TiTok is also effective in the personalization setting. Among the LaMP tasks, we focus on the two text generation tasks that are suitable, accessible, and evaluable in our setting:

- **News Headline Generation (LaMP 4).** Given an author profile consisting of previously written headlines, the model is asked to generate a headline for a new news article. The task evaluates if the model can adapt its output to reflect the author’s characteristic style in journalistic writing.
- **Scholarly Title Generation (LaMP 5).** Using an author profile built from prior titles of academic publications, the model generates a title for a new given abstract. The task assesses the model’s ability to capture and reproduce distinctive conventions of academic writing.

864
865
866 Table 7: Categorization of the 27 Big-Bench Hard (BBH) tasks.
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
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

Category	Tasks
Logical Reasoning	boolean_expressions, causal_judgement, date_understanding, disambiguation_qa, dyck_languages, formal_fallacies, logical_deduction_three_objects, logical_deduction_five_objects, logical_deduction_seven_objects, temporal_sequences, tracking_shuffled_objects_three_objects, tracking_shuffled_objects_five_objects, tracking_shuffled_objects_seven_objects, web_of_lies
Linguistic	hyperbaton, ruin_names, salient_translation_error_detection, snarks, word_sorting
Mathematical / Symbolic	geometric_shapes, multistep_arithmetic_two, object_counting, reasoning_about_colored_objects
Applied / Knowledge	movie_recommendation, navigate, penguins_in_a_table, sports_understanding

879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
880 Table 8: Categorization of the 57 MMLU tasks.
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
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

Category	Tasks
STEM (29)	abstract_algebra, anatomy, astronomy, college_biology, college_chemistry, college_computer_science, college_mathematics, college_medicine, college_physics, computer_security, conceptual_physics, electrical_engineering, elementary_mathematics, formal_logic, high_school_biology, high_school_chemistry, high_school_computer_science, high_school_mathematics, high_school_physics, high_school_statistics, machine_learning, medical_genetics, nutrition, professional_medicine, professional_psychology, virology, clinical_knowledge, human_aging, human_sexuality
Humanities (16)	business_ethics, formal_fallacies, jurisprudence, logical_fallacies, philosophy, prehistory, world_religions, moral_disputes, moral_scenarios, professional_law, professional_accounting, high_school_world_history, high_school_us_history, high_school_european_history, global_facts, security_studies
Social Sciences (10)	econometrics, high_school_macroeconomics, high_school_microeconomics, management, marketing, public_relations, sociology, us_foreign_policy, high_school_geography, high_school_government_and_politics
Other (2)	miscellaneous, global_facts

D BASELINE DETAILS

903 To evaluate the effectiveness of TiTOK, we compare against several baselines as follows:

D.1 VANILLA

907 The vanilla baseline corresponds to the standard target base model without any additional training or
908 knowledge transfer from the source model. This setup reflects the performance of the target model
909 in its raw initialized state and serves as a lower bound for evaluating transfer methods. Intuitively,
910 achieving performance that surpasses this baseline provides clear evidence that the target model has
911 successfully acquired and internalized knowledge transferred from the source model.

D.2 KNOWLEDGE DISTILLATION (KD) (+MINED)

915 The knowledge distillation (KD) (Hinton et al., 2015; Azimi et al., 2024) baseline trains the student
916 model to mimic the teacher model’s output distribution. Specifically, the loss is a weighted sum of
917 the cross entropy objective and the KL divergence between the teacher and student distributions. In
918 our setup, the KD experiments are conducted using the TransLoRA filtered synthetic datasets.

918 For cases where the source and target models use mismatched tokenizers when performing KD,
 919 we apply the Minimal Edit Distance (MinED) alignment method (Wan et al., 2024). In particular,
 920 MinED matches tokens across different vocabularies by minimizing the number of character
 921 level edits. For example, it can align “get” with “gets,” “color” with “colour,” or “analysis” with
 922 “analyses.” This tokenizer alignment approach avoids degeneracy issues in token alignment when
 923 conducting KD, though it is different from our dual-pointer based alignment algorithm.

925 D.3 TRANSLoRA

927 The TransLoRA baseline (Wang et al., 2024) transfers LoRA knowledge by generating synthetic
 928 data. In this method, the vanilla target model synthesizes queries, and the source model with its
 929 LoRA adapter provides the corresponding labels. Subsequently, a discriminator, separately trained
 930 with the source model as its base, is applied to filter the synthetic data used for fine-tuning the
 931 target model’s LoRA adapter. For the sake of consistency and fair comparison, we adopt the same
 932 hyperparameter settings from our experiments when applying the TransLoRA baseline.

933 **Comparison with TransLoRA.** TransLoRA’s core assumption is to transfer entire synthetic
 934 sequences generated by the teacher. Therefore, its framework is tightly coupled to synthetic data and
 935 is unable to function in the absence of teacher-generated labels. In that sense, this also implies that
 936 TransLoRA cannot be applied to external datasets. Thus, it mainly highlights the role of synthetic
 937 data, giving less thought to the architecture of the knowledge transfer procedure itself.

938 In contrast, TiTOK directly rethinks the mechanism of knowledge transfer itself. TiTOK
 939 concentrates on determining which tokens actually contain expert-specific knowledge (i.e. the tokens
 940 where the expert significantly deviates from the basic model) rather than copying entire synthetic
 941 sequences. Our novel token-level contrastive signal fundamentally changes how knowledge is ex-
 942 tracted and aligned. Because this process does not rely on teacher-generated labels, TiTOK naturally
 943 extends to external, non-synthetic datasets (See Table 3) and is therefore much more generalizable.

944 Therefore, despite the fact that the ultimate objective of both approaches is to solve the same
 945 high-level LoRA transfer problem, which inevitably leads to certain overlapping components, the
 946 underlying philosophies are essentially different. TransLoRA focuses on transferring full outputs
 947 and relies entirely on synthetic data, whereas TiTOK focuses on transferring the expert’s internal
 948 knowledge signals regardless of the data source, thus being more generalizable and effective.

950 E USAGE OF AI ASSISTANTS

951 AI assistants are minimally used in this work, restricted solely to language refinement, such as
 952 grammar correction, punctuation, and sentence structure. All research ideas, methodologies, and
 953 analyses are the original contributions of the authors. Thus, the use of AI is confined to editorial
 954 support and did not influence the originality or intellectual contributions of the work.

957 F EXTENDED TABLE 1 RESULTS WITH VARIANCE AND STATISTICAL DETAIL

958 We additionally report the mean \pm standard deviation over 3 random seeds of our results in Table
 959 1. The results are presented in Table 9. As shown, the overall trend remains consistent: relative to
 960 the target vanilla baseline, KD improves by 1.25%, TransLoRA by 5.28%, and TiTOK by 9.94%.
 961 In fact, these improvements remain substantially larger than the corresponding standard deviations

966 G COMPARISON WITH ADDITIONAL BASELINE

967 Here, we introduce an additional baseline that leverages few-shot supervision. In practice, our
 968 synthetic data generation process is seeded with a small set of prompts, which can be regarded
 969 as few-shot data. To be thorough, we therefore establish a baseline trained only on these five
 970 seed prompts, referred to as KD (5-shot), so that our evaluation of TiTOK also considers minimal
 971 few-shot supervision. The results of this additional baseline are presented in Table 10.

972 Table 9: **Main results with variance.** Mean \pm standard deviation over three random seeds on BBH,
 973 MMLU, News Headline, and Scholarly Title under four transfer settings. BBH and MMLU are eval-
 974 uated by exact-match accuracy, while News Headline and Scholarly Title are evaluated by ROUGE-
 975 1/L. All evaluations are zero-shot. Best scores are in **bold**, while second highest are underlined.

Transfer	Method	BBH		MMLU		News Headline		Scholarly Title	
		Acc.	Acc.	Acc.	R-1	R-L	R-1	R-L	
Mistral 7B → Mistral 7B	Vanilla	0.397	0.557	0.117	0.101	0.381	0.311		
	KD	<u>0.417</u> \pm 0.007	<u>0.560</u> \pm 0.003	0.117 \pm 0.004	0.104 \pm 0.003	0.385 \pm 0.006	0.310 \pm 0.005		
	TransLoRA	0.416 \pm 0.006	0.534 \pm 0.001	0.156 \pm 0.002	<u>0.137</u> \pm 0.001	<u>0.447</u> \pm 0.001	<u>0.382</u> \pm 0.001		
	TiTOK (ours, k=70%)	0.424 \pm 0.008	0.561 \pm 0.002	0.161 \pm 0.001	0.143 \pm 0.001	0.473 \pm 0	0.413 \pm 0.001		
Mistral 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378		
	KD	<u>0.475</u> \pm 0.004	<u>0.482</u> \pm 0.004	<u>0.127</u> \pm 0	<u>0.112</u> \pm 0.001	<u>0.454</u> \pm 0.001	<u>0.387</u> \pm 0.002		
	TransLoRA	0.473 \pm 0.003	0.473 \pm 0.001	0.126 \pm 0.003	0.110 \pm 0.002	<u>0.461</u> \pm 0.001	<u>0.397</u> \pm 0.001		
	TiTOK (ours, k=70%)	0.484 \pm 0.002	0.485 \pm 0.003	0.139 \pm 0.001	0.123 \pm 0.001	0.464 \pm 0.001	0.403 \pm 0.001		
Llama3 3B → Llama3 8B	Vanilla	0.469	0.469	<u>0.125</u>	<u>0.110</u>	0.444	0.378		
	KD	<u>0.474</u> \pm 0.005	<u>0.477</u> \pm 0.002	<u>0.125</u> \pm 0.001	<u>0.110</u> \pm 0.001	<u>0.449</u> \pm 0.001	<u>0.383</u> \pm 0.001		
	TransLoRA	0.471 \pm 0.009	0.467 \pm 0.002	0.122 \pm 0.001	0.108 \pm 0.001	<u>0.454</u> \pm 0.001	<u>0.387</u> \pm 0.001		
	TiTOK (ours, k=30%)	0.496 \pm 0.011	0.478 \pm 0.004	0.127 \pm 0.001	0.113 \pm 0.001	0.456 \pm 0.001	0.392 \pm 0		
Llama2 7B → Llama3 8B	Vanilla	0.469	0.469	<u>0.125</u>	<u>0.110</u>	0.444	0.378		
	KD	<u>0.473</u> \pm 0.002	<u>0.476</u> \pm 0.002	<u>0.125</u> \pm 0	<u>0.110</u> \pm 0.001	<u>0.449</u> \pm 0.001	<u>0.382</u> \pm 0.002		
	TransLoRA	0.472 \pm 0.002	0.468 \pm 0.002	0.123 \pm 0.001	0.109 \pm 0.001	<u>0.457</u> \pm 0.003	<u>0.394</u> \pm 0.005		
	TiTOK (ours, k=70%)	0.488 \pm 0.019	0.477 \pm 0.003	0.138 \pm 0.002	0.120 \pm 0.002	0.461 \pm 0.001	0.403 \pm 0.001		

993 Table 10: **Comparison with a few-shot KD baseline using five seed prompts.** BBH and MMLU
 994 are reported as the average accuracy across tasks, while News Headline and Scholarly Title Gener-
 995 ation tasks are evaluated using ROUGE-1 (R-1) and ROUGE-L (R-L). *KD (5-shot)* denotes knowl-
 996 edge distillation performed with only five few-shot examples. Best scores are in **bold**.

Transfer	Method	BBH		MMLU		News Headline		Scholarly Title	
		Acc.	Acc.	Acc.	R-1	R-L	R-1	R-L	
Mistral 7B → Mistral 7B	Vanilla	0.397	0.557	0.117	0.101	0.381	0.311		
	KD (5-shot)	0.402	0.558	0.118	0.104	0.383	0.312		
	TiTOK (ours)	0.432	0.563	0.160	0.142	0.473	0.414		
Mistral 7B → Llama 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378		
	KD (5-shot)	0.470	0.477	0.126	0.111	0.446	0.379		
	TiTOK (ours)	0.482	0.488	0.140	0.124	0.464	0.403		
Llama3 3B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378		
	KD (5-shot)	0.470	0.479	0.126	0.111	0.446	0.379		
	TiTOK (ours)	0.509	0.475	0.127	0.113	0.457	0.392		
Llama2 7B → Llama3 8B	Vanilla	0.469	0.469	0.125	0.110	0.444	0.378		
	KD (5-shot)	0.470	0.477	0.126	0.111	0.446	0.380		
	TiTOK (ours)	0.510	0.479	0.140	0.122	0.461	0.404		

1017 Across all transfer settings, KD on only 5-shot samples provides only marginal improvements over
 1018 the vanilla model, with average gains of 0.9% on reasoning tasks (BBH and MMLU) and 0.6–1.2%
 1019 on personalization tasks (News Headline and Scholarly Title Generation). In contrast, TiTOK
 1020 delivers substantial improvements over this baseline, achieving 4.7% and 2.1% average gains on
 1021 reasoning tasks and 8.9–16.6% average gains on personalization tasks. Winning over KD with only
 1022 5 few-shot samples shows that TiTOK is not simply the result of a few-shot effect. Instead, the
 1023 five seed prompts act only as a starting point, which our framework expands into richer and more
 1024 effective synthetic training signals. This finding confirms that the true strength of TiTOK lies in
 1025 how it effectively leverages limited supervision rather than in the few-shot data itself.

1026 Table 11: TiTOK performance across $k = 10\%-90\%$. “Vanilla” denotes performance of target
 1027 vanilla model. Highlighted columns denote the selected universal $k\%$ reported in Table 1. Highest
 1028 performance are in **bold**, while the second best results are underlined.

Transfer	Task	Metric	Vanilla	10%	20%	30%	40%	50%	60%	70%	80%	90%	
Mistral 7B → Mistral 7B	News Headline	BBH	Acc	0.397	0.401	0.420	<u>0.444</u>	0.445	0.441	0.432	<u>0.432</u>	0.431	0.431
		MMLU	Acc	0.557	0.556	0.556	0.558	0.553	0.554	<u>0.560</u>	0.563	<u>0.560</u>	0.558
		R-1	0.117	0.153	0.159	0.161	0.161	<u>0.160</u>	<u>0.160</u>	<u>0.160</u>	<u>0.160</u>	<u>0.160</u>	0.161
		R-L	0.101	0.138	<u>0.143</u>	0.144	<u>0.143</u>	0.142	0.142	0.142	0.142	0.142	0.142
		Scholarly Title	R-1	0.381	0.466	<u>0.475</u>	0.481	0.474	0.473	0.473	0.473	0.473	0.470
		R-L	0.311	0.408	<u>0.419</u>	0.424	0.416	0.414	0.414	<u>0.414</u>	0.414	0.414	0.411
	News Headline	BBH	Acc	0.469	0.470	<u>0.482</u>	0.471	0.473	0.475	0.476	<u>0.482</u>	0.483	0.478
		MMLU	Acc	0.469	0.487	0.483	0.500	0.492	0.492	0.488	0.488	<u>0.494</u>	0.500
		R-1	0.125	0.140	<u>0.141</u>	0.142	0.142	<u>0.141</u>	0.140	0.140	0.138	0.138	0.138
		R-L	0.110	0.123	0.124	<u>0.125</u>	0.126	0.124	0.123	0.124	0.122	0.122	0.121
		Scholarly Title	R-1	0.444	0.460	0.458	0.458	0.460	0.467	<u>0.466</u>	0.464	0.465	0.465
		R-L	0.378	0.398	0.394	0.395	0.396	0.406	<u>0.405</u>	0.403	0.403	0.406	
Llama3 3B → Llama3 8B	News Headline	BBH	Acc	0.469	<u>0.515</u>	0.507	<u>0.509</u>	0.512	0.500	0.492	0.505	0.518	0.509
		MMLU	Acc	0.469	0.479	<u>0.477</u>	0.475	0.475	0.475	0.474	0.475	0.474	0.472
		R-1	0.125	<u>0.127</u>	<u>0.127</u>	<u>0.127</u>	0.128	0.122	0.121	0.122	0.121	0.121	0.121
		R-L	0.110	0.111	<u>0.112</u>	0.113	<u>0.113</u>	0.108	0.107	0.107	0.107	0.107	0.107
		Scholarly Title	R-1	0.444	0.449	0.456	0.457	0.460	0.462	<u>0.462</u>	0.461	0.462	0.462
		R-L	0.378	0.385	0.390	<u>0.392</u>	0.397	0.400	<u>0.400</u>	0.398	0.398	0.398	0.397
	News Headline	BBH	Acc	0.469	<u>0.527</u>	0.519	0.531	0.510	0.514	0.514	<u>0.510</u>	0.509	0.509
		MMLU	Acc	0.469	0.475	0.474	0.477	<u>0.481</u>	0.477	0.482	0.479	0.477	0.479
		R-1	0.125	0.126	0.129	0.130	0.135	0.138	0.138	<u>0.140</u>	<u>0.139</u>	<u>0.139</u>	
		R-L	0.110	0.110	0.114	0.115	0.119	<u>0.121</u>	0.120	0.122	<u>0.121</u>	<u>0.121</u>	
		Scholarly Title	R-1	0.444	0.450	0.455	0.453	0.453	0.458	0.460	<u>0.461</u>	<u>0.461</u>	0.464
		R-L	0.378	0.385	0.391	0.389	0.389	0.396	0.402	<u>0.404</u>	<u>0.404</u>	0.406	

H COMPREHENSIVE K% SENSITIVITY ANALYSIS

In this section, we conduct an extensive experiment on TiTOK’s performance throughout a wide range of k values, from 10% to 90%. The results are comprehensively presented in Table 11. These experiments demonstrate that TiTOK consistently outperforms the target vanilla baseline throughout a broad, steady range of values, while the exact performance varies slightly depending on k%. This clearly indicates that the method is not overly sensitive to the choice of k% and that there exists a broad range of reasonable k% values where improvements are reliably obtained.

The reason for choosing a universal k% hyperparameter in our study is to keep the presentation of our paper more coherent and consistent. For this reason, we intentionally avoid task-specific tuning and report the k% that generally works well across tasks. However, we note that with additional per-task optimization of k%, further performance improvements are indeed possible.

With regard to the adaptive mechanism, each transfer setting exhibits its own effective k% range. For instance, same-backbone BBH transfer favors a mid-range k% (40–70%), weak-to-strong BBH transfer benefits from smaller k%, and weak-to-strong News Headline Generation shows the opposite trend, preferring larger k%. Taken together, these trends suggest that a simple adaptive mechanism can already be practical and effective. In our experiments, even coarse adjustments (i.e using around 30% for weak-to-strong reasoning transfers and 70% for weak-to-strong stylistic tasks as denoted in Section 4.3) prove sufficient, without requiring extensive hyperparameter searches.

I COMPUTATIONAL OVERHEAD COMPARISON

Here, we present the computational differences between TiTOK and TransLoRA. We conduct an end-to-end time comparison of TransLoRA and TiTOK on the BBH and News Headline Generation benchmarks in the Mistral 7B → Llama3 8B transfer setting. The results are presented in Table 12 and 13. Across both benchmarks, TiTOK achieves roughly a 1.5x–2.5x reduction averagely in

BBH (250 data × 27 tasks)					
Method	Metric	discriminator training	selecting data	training	TOTAL
TransLoRA	avg_total (sec/num.task)	279.33936	31.72915	104.61636	415.68487
	avg_sample (sec/sample)	0.63251	0.06346	0.41847	1.11444
TiTok	avg_total (sec/num.task)	164.70895	0.01155	0.41080	102.02867
	avg_sample (sec/sample)	0.32942	0.00005	0.00054	0.40811

Table 12: Runtime comparison for Mistral 7B → Llama3 8B transfer on BBH.

News Headline (200 data × 30 users)					
Method	Metric	discriminator training	selecting data	training	TOTAL
TransLoRA	avg_total (sec/num.users)	245.10085	25.39730	58.71148	329.20963
	avg_sample (sec/sample)	0.61279	0.06358	0.29356	0.96993
TiTok	avg_total (sec/num.users)	81.31437	0.01863	0.89304	49.52768
	avg_sample (sec/sample)	0.20355	0.00005	0.00447	0.24764

Table 13: Runtime comparison for Mistral 7B → Llama3 8B transfer on News Headline Generation.

total compute time. Most of the gain comes from the removal of discriminator training, while the per-token log-likelihood computation produces only a minimal and manageable overhead.

J ADDITIONAL EXPERIMENTS ON LARGER-SCALE MODELS AND ARCHITECTURES

While we have already considered transfer settings across various model sizes and architectures, we additionally conduct experiments on larger-scale models and architectures to reinforce TiTok’s effectiveness even in these expanded settings. Note that all experiments are executed under 4-bit quantization due to the large model sizes involved. We also evaluate in BBH for this whole section.

First, we run experiments on cross-architecture transfer from a dense model (Mistral 7B) to a mixture-of-experts (MoE) model (Mixtral-8×7B-Instruct-v0.1). The results are presented in Table 14. Analytically, TiTok outperforms the MoE target baseline, indicating that TiTok remains effective in the MoE setting. This result suggests that TiTok can generalize to heterogeneous model architectures even when transferring from a dense model to a MoE model.

Furthermore, we test also in the Llama-3.3-70B-Instruct → Llama-3.3-70B-Instruct transfer setting. As shown in Table 15, TiTok effectively transfers knowledge even in this 70B transfer setting, with 4.7% improvement. This result shows that TiTok generalizes well to very large models, preserving its ability to transfer token-level knowledge effectively.

K SYNTHETIC DATA GENERATION WITH 2× POOL AND TOP- M SELECTION

In this section, we provide further details regarding the number of synthetic data samples used. In generating synthetic data, we first provide five samples from the original training set as few-shot exemplars to guide the model toward producing outputs in the desired style and format. For each task, we initially create a synthetic pool containing twice the number of examples used in the source model’s training. This pool is then filtered using token-wise contrastive excess scores, after which we retain only the top M samples, where M equals the size of the source training set. To be specific, we set $M = 250$ for BBH, $M = 90$ for MMLU, and $M = 200$ for LaMP tasks. This procedure ensures that the target model’s LoRA adapter is trained on a dataset comparable in scale to that of the source model, while selective filtering enhances the overall quality of the retained data.

1134	Model Transfer	Metric	Acc.
1135	Mistral 7B	Vanilla	0.454
1136	→ Mixtral-8×7B-Instruct-v0.1	TiTOK (k=70%)	0.463
1137			
1138			

1139 Table 14: BBH results for the Mistral 7B → Mixtral-8×7B-Instruct-v0.1 transfer setting.
1140 **TiTOK (k=70%)** yields an improvement of approximately 2% over the target vanilla model.

1142	Model Transfer	Metric	Acc.
1143	Llama-3.3-70B-Instruct	Vanilla	0.593
1144	→ Llama-3.3-70B-Instruct	TiTOK (k=70%)	0.621
1145			
1146			

1147 Table 15: BBH results for the Llama-3.3-70B-Instruct → Llama-3.3-70B-Instruct
1148 transfer setting. **TiTOK (k=70%)** improves over the target vanilla model.

L ROUGE-L FILTERING FOR DIVERSE SYNTHETIC QUERIES

1153 We now proceed to provide the task lists for which we did not apply ROUGE-L filtering when
1154 generating diverse synthetic queries. In general, we use ROUGE-L filtering to encourage diversity
1155 in queries, but for the tasks listed in Table 16, we only applied simple deduplication. In the case of
1156 BBH, tasks such as *boolean_expressions* or *temporal_sequences* already follow highly restricted and
1157 repetitive patterns, making high ROUGE-L scores inevitable. For MMLU, the only tasks without
1158 ROUGE-L filtering are the history-related subjects. This is potentially because history questions
1159 often require long passages that overlap in vocabulary, phrasing, or factual references (e.g., recurring
1160 names, dates, or events). Applying strict ROUGE-L filtering in such cases would make it difficult
1161 to generate the required number of synthetic queries. For all remaining BBH and MMLU tasks, as
1162 well as all LaMP tasks, we applied a ROUGE-L threshold of 0.7 to encourage diversity while still
1163 preserving task fidelity. Table 16 provides the tasks for which ROUGE-L filtering is not applied.

1164 Table 16: Tasks without ROUGE-L filtering.

1165 Category	1166 Tasks
1167 BBH	bbh_boolean_expressions, bbh_date_understanding, 1168 bbh_disambiguation_qa, bbh_geometric_shapes, 1169 bbh_logical_deduction_three_objects, bbh_multistep_arithmetic_two, 1170 bbh_navigate, bbh_object_counting, bbh_penguins_in_a_table, 1171 bbh_reasoning_about_colored_objects, 1172 bbh_salient_translation_error_detection, bbh_snarks, 1173 bbh_temporal_sequences, bbh_tracking_shuffled_objects_three_objects, 1174 bbh_web_of_lies
1175 MMLU	high_school_world_history, high_school_us_history, 1176 high_school_european_history

M ROBUSTNESS OF TiTOK TO SYNTHETIC DATA QUALITY

1178 While TiTOK’s primary contribution is token-level transfer and is, thus, not tied to synthetic data,
1179 we provide a detailed analysis on the robustness of our method to synthetic-data variations.

1180 **Robustness to low-quality synthetic data.** First, we deliberately construct a suboptimal
1181 synthetic dataset by selecting only the 250 lowest-scoring synthetic samples, as evaluated by the
1182 GPT-4o-mini grading model and prompt in (scores ∈ {0, 1, 2, 3, 4, 5}) (Chen et al., 2024a). All
1183 training hyperparameters and filtering processes are identical to the original setup. As shown in
1184 Table 17, TiTOK still substantially outperforms all baselines in the Mistral 7B → Mistral 7B BBH
1185 transfer setting. This shows that TiTOK is still effective even with low-quality synthetic data.

Model Transfer	Method	BBH (Acc.)
	Vanilla	0.397
Mistral 7B → Mistral 7B	KD	0.406
	TransLoRA	0.405
	TiTOK (k=70%)	0.416

Table 17: BBH exact-match performance for the Mistral 7B → Mistral 7B transfer setting using low-quality synthetic data. TiTOK (k=70%) continues to outperform all baselines.

Table 18: Alternative synthetic query prompt template for BBH.

Synthetic query generation prompt for {task.name} in BBH.

Generate {task.name} questions like these examples:

Example 1:
(few-shot example 1)

Example 2:
(few-shot example 2)

Example 3:
(few-shot example 3)

Example 4:
(few-shot example 4)

Example 5:
(few-shot example 5)

Example 6:

Robustness to data diversity. Our synthetic data generation pipeline already incorporates deduplication and a ROUGE-L diversity threshold. This ensures the exclusion of exact duplicates and highly similar samples. More details are provided in Appendix L.

Robustness to prompt design. We further conduct an experiment with a different prompt for generating synthetic data. The alternative prompt template is provided in Table 18. We test this on BBH tasks in the Mistral 7B → Mistral 7B and Mistral 7B → Llama3 8B transfer settings.

Table 19 shows the results of using the alternative prompt for generating synthetic data. Analytically, the results show that TiTOK achieves consistent improvements, and there are no substantial differences across prompt choices. This demonstrates that TiTOK is not sensitive to prompt formulation, as long as the generated examples adhere to a reasonable and coherent structural pattern.

Overall, we emphasize that a highly refined or meticulously selected synthetic dataset is not required for TiTOK. As long as the generated samples retain a relatively cohesive task structure, TiTOK is still effective and is not very sensitive to the particular features of the synthetic data.

1242 Table 19: Effect of prompt variation on TiTOK on BBH in the Mistral 7B → Mistral 7B and Mistral
1243 7B → Llama3 8B settings.
1244

1245 Transfer	1246 Method	1247 Prompt	1248 BBH (Acc.)
1247 Mistral 7B → Mistral 7B	1248 TiTOK (k=70%)	Vanilla	1249 – 0.397
		original (Table 23)	1250 0.432
		alternative (Table 18)	1251 0.436
1250 Mistral 7B → Llama3 8B	1251 TiTOK (k=70%)	Vanilla	1252 – 0.469
		original (Table 23)	1253 0.482
		alternative (Table 18)	0.485
1254	1255	1256	1257
1258 Data Source	1259 BBH	1260 Scholarly Title	1261
1262 Gold data	1263 2.40	3.68	1264
Mistral 7B synthetic data	2.27	3.89	1265
Llama3 3B synthetic data	2.01	3.74	1266
Llama2 7B synthetic data	2.07	3.15	1267

1262 Table 20: Overall average quality ratings of gold and synthetic datasets across the BBH and Scholarly
1263 Title Generation tasks. GPT-4o-mini evaluates each data sample on a 1–5 scale.
12641265

N ADDRESSING SYNTHETIC DATA QUALITY AND BIAS

1266

1267 We now move on to explain how our pipeline incorporates several components specifically designed
1268 to reduce bias and prevent quality degradation in synthetic data. Specifically, we achieve this by in-
1269 corporating several components, including a length filter that removes malformed or low-informative
1270 samples, a token-wise token-wise contrastive excess mechanism that selects only the most informa-
1271 tive tokens rather than relying on entire synthetic sentences, and diversity filtering (i.e deduplication
1272 and ROUGE-L thresholding) to prevent mode collapse during synthetic data generation.

1273 Additionally, we further execute an external quality evaluation using a separate model
1274 (GPT-4o-mini), adopting the robust synthetic-data assessment methodology established in prior
1275 work (Chen et al., 2024a). Table 20 reports the average quality scores (scale 1–5) of gold data and
1276 the synthetic data generated by multiple source models used in our experiments.

1277 These results show that the synthetic data used in our experiments is comparable in quality to the
1278 gold dataset, with no evidence of harmful bias dominating the data. Importantly, TiTOK performs
1279 consistently well across all synthetic sources, even when the quality varies (e.g., 2.01 vs. 2.40 in
1280 BBH), which further indicates that TiTOK does not rely on perfectly clean or unbiased synthetic
1281 data. Taken together, these findings demonstrate that TiTOK is robust to imperfections in synthetic
1282 data, and, once again, can be effectively used even in non-synthetic datasets as shown in Table 3.
1283

1284

O ROBUSTNESS OF TiTOK UNDER WEAKER SOURCE LORAS

1285

1286 We now discuss scenarios where the source LoRA is weaker than the target model. While it is true
1287 that TiTOK uses the source LoRA as a reference, its effectiveness is not strongly dependent on the
1288 source LoRA’s absolute performance. To illustrate this point more clearly, Table 21 summarizes the
1289 subsets of Table 1 in which the source LoRA underperforms the target model’s vanilla version, and
1290 presents these source LoRA scores alongside the corresponding transfer results.
1291

1292 As shown in the table, TiTOK still consistently improves the target model even when the source
1293 LoRA is weaker than the target baseline. This shows that TiTOK captures domain-specific signals
1294 that are embedded in key tokens rather than inheriting the source LoRA’s full behavior.
1295

1296	1297	1298	Model Transfer	Metric	BBH		Scholarly Title	
					1299	1300	1301	1302
1300	1301	1302	Llama3 3b → Llama3 8b	source lora	0.460	0.429	0.363	
				target base	0.462	0.444	0.378	
				TiTOK (k=30%)	0.509	0.457	0.392	
1303	1304	1305	Llama2 7B → Llama3 8B	source lora	0.359	0.431	0.367	
				target base	0.469	0.444	0.378	
				TiTOK (k=70%)	0.510	0.461	0.404	

Table 21: Results on BBH and Scholarly Title Generation for transfer settings where the source LoRA is weaker than the target baseline.

Table 22: “Initial TiTOK” indicates TiTOK applied once. “Iterative TiTOK: *self transfer*” applies TiTOK again using synthetic data generated by the learned target LoRA after the initial transfer.

1312			1313		BBH (Acc.)		
1314			Mistral 7B → Mistral 7B			Vanilla	0.397
						Initial TiTOK	0.432
						Iterative TiTOK: <i>self transfer</i>	0.456
						Target expert	0.460
1319			Mistral 7B → Llama3 8B			Vanilla	0.469
						Initial TiTOK	0.482
						Iterative TiTOK: <i>self transfer</i>	0.510
						Target expert	0.531

P EXPLORATION ON SELF REFINEMENT

While TiTOK primarily focuses on LoRA-to-LoRA knowledge transfer, we explore an intriguing new direction by examining whether TiTok can also support self-refinement. We perform this experiment on the BBH benchmark in the Mistral 7B → Mistral 7B and the Mistral 7B → Llama3 8B transfer settings. In particular, after the initial transfer step, 1) we generate synthetic data using the target LoRA that had already learned TiTOK’s transferred knowledge, 2) compute contrastive excess scores on the newly generated synthetic data using the original source LoRA and the its base model, and 3) train a new target LoRA with the same hyperparameters in the original setting (e.g., k=0.7%). The empirical results for self-refinement are summarized above in Table 22.

Interestingly, the results show that the iterative step yields additional improvement, implying that iterative refining is possible. We hypothesize that this is possible because the initial transfer increases the target model’s familiarity with the task and partially aligns its internal representations with the source domain. As a result, the target model generates cleaner, more coherent synthetic data with fewer irrelevant patterns. This thereby enables the second iteration of TiTOK to more accurately identify domain-informative tokens, which improves token selection and makes excess scoring more discriminative. However, we emphasize that iterative refining synthetic data cannot surpass a target model trained on real train data. By creating increasingly better synthetic data, iteration can help narrow the gap, but it is still far from a fully supervised target model trained on actual data.

Q QUALITATIVE EXAMPLES

In this section, we provide qualitative examples from the News Headline Generation in the Mistral 7B → Mistral 7B transfer setting to illustrate how TiTOK identifies the most informative tokens. The qualitative examples are presented in Figure 5. Analytically, the selected tokens correspond to

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

Prompt	<p>You are a news headline generator.</p> <p>Generate a headline for the following article.</p> <p>article: With a scalloped hem, it's a show-stopping piece that's both unique and affordable. Want more? Be sure to check out Stylelist on....</p> <p>headline:</p>
Output	<p>Stylist: The Perfect White Dress For Your Wedding (PHOTOS) </s></p>
Prompt	<p>You are a news headline generator.</p> <p>Generate a headline for the following article.</p> <p>article: It's hard to imagine that you would be interested in hearing about my favorite holiday tradition, but if you are, I'd be happy to share....</p> <p>headline:</p>
Output	<p>How To Make Your Own DIY Christmas Ornaments (PHOTOS) </s></p>

Figure 5: **Qualitative examples from News Headline Generation in the Mistral 7B → Mistral 7B setting.** **Bold** and **highlighted** tokens are part of the selected tokens. Notably, these tokens exhibit high token-wise contrastive excess scores and thus fall into the top 70% selected for training.

the beginnings of major semantic or structural units in the headline. These tokens carry the highest information value: they denote transitions, introduce core noun phrase components, or contain root words. Moreover, since the token-wise contrastive excess score is defined as the difference between the expert model (source LoRA + base) and the amateur model (base only), it pinpoints the tokens where the expert and amateur diverge most in their predictions and so include effective learning signals for target tasks. Taken together, these qualitative examples effectively demonstrate that TiTOK consistently retains the tokens that are most important for providing structure and meaning.

KD-based approaches, however, are highly dependent on the absolute capacity of the source LoRA. By making the target model follow the teacher’s logits, a weak or poorly tuned source LoRA unavoidably transfers certain flaws, resulting in the minimal gains supported by the results in the above table. Meanwhile, TiTOK avoids these deficiencies because it fundamentally differs in how it leverages the source LoRA. Instead of relying on the source LoRA’s logits globally, TiTOK uses the source only to smartly identify informative token positions. As a result, a poorly tuned source LoRA does not mislead TiTOK since it does not make the target model imitate the source LoRA’s behavior. Rather, it selectively retrieves the useful tokens using the source LoRA’s domain knowledge.

R PROMPTS FOR SYNTHETIC QUERY GENERATION

In this section, we present the prompts used for generating synthetic queries. Each prompt includes five examples, which correspond to the seed prompts taken from the original data. For BBH, which consists of multiple subtasks, we specify task-specific instructions and formatting rules so that the generated queries and labels are structured according to the expected format. Each BBH subtask is paired with its corresponding rules and output format to ensure consistency across the benchmark, and Table 23 illustrates the case of *boolean_expressions* as a representative example.

In contrast, for MMLU a single unified prompt format is sufficient for most subjects, and the model reliably generates queries that follow the expected structure. The exception is the history-related tasks, namely, *high_school_us_history*, *high_school_world_history*, and

1404 Table 23: Synthetic query prompt for the BBH *boolean_expressions* task. Each BBH subtask re-
 1405 quires task-specific instructions and formatting rules to ensure that generated queries and labels
 1406 follow the expected format. We show *boolean_expressions* task here as a representative example.
 1407

1408 **Synthetic query generation prompt for BBH (*boolean_expressions*) task.**

1409 **System**

1410 You are an expert task generator for boolean_expressions tasks. Generate boolean expression
 1411 evaluation tasks ending with ' is ' – SINGLE LINE ONLY.
 1412

1413 **CRITICAL FORMAT REQUIREMENTS:**

- 1414 - Follow the EXACT format structure shown in the examples.
- 1415 - ABSOLUTELY CRITICAL: Generate **ONLY ONE LINE**, ONE TASK per response.
- 1416 - NO multiple lines, NO newlines (\n), NO multiple expressions.
- 1417 - Must end with ' is '.
- 1418 - Example of **INVALID** output: True or False is\n\n False and True is
- 1419 - Example of **VALID** output: True or False or not False and (True or
 1420 False) is

1421 Generate diverse content but maintain the exact same format structure. Only output the task
 1422 input, not the solution.
 1423

1424 **User**

1425 Generate new boolean_expressions tasks following these exact format examples:

- 1426 Example 1: [boolean expression ending with ' is ']
- 1427 Example 2: [boolean expression ending with ' is ']
- 1428 Example 3: [boolean expression ending with ' is ']
- 1429 Example 4: [boolean expression ending with ' is ']
- 1430 Example 5: [boolean expression ending with ' is ']

1431 CRITICAL: Generate ONLY ONE LINE, exactly like the examples above.

1432 Generate a new task following the exact format:

1435 *high_school_european_history*. For these tasks, it is observed that more specific prompting is
 1436 required to obtain better generations. The general MMLU prompt is shown in Table 24, whereas
 1437 the history-related tasks make use of the specialized prompt provided in Table 25.

1438 Finally, for LaMP tasks, we apply the same query template uniformly to each user. The detailed
 1439 prompt templates for LaMP 4 and 5 are provided in Tables 26 and 27 respectively.

1441
 1442
 1443
 1444
 1445
 1446
 1447
 1448
 1449
 1450
 1451
 1452
 1453
 1454
 1455
 1456
 1457

1458

1459

1460

1461

1462

1463 Table 24: Synthetic query prompt for the MMLU. Applied to all tasks except for history related
1464 tasks.

1465

1466 **Synthetic query generation prompt for MMLU (general).**

1467

Example 1:

Question: [question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1474

Example 2:

Question: [question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1482

Example 3:

Question: [question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1489

Example 4:

Question: [question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1497

Example 5:

Question: [question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1505

Example 6:

1508

1509

1510

1511

1512

1513

1514

1515 Table 25: Synthetic query prompt for the MMLU history tasks only. The history tasks are
 1516 *high_school_us_history*, *high_school_world_history*, and *high_school_european_history*. These his-
 1517 tory tasks required more specific prompting to obtain better results.

1518

1519

1520

Synthetic query generation prompt for MMLU (history related tasks).

1521

Generate [SUBJECT] multiple choice questions following these examples:

1522

Example 1:

Question: [history question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1529

Example 2:

Question: [history question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1537

Example 3:

Question: [history question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1544

Example 4:

Question: [history question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1552

Example 5:

Question: [history question snippet]

1. [choice snippet]

2. [choice snippet]

3. [choice snippet]

4. [choice snippet]

Answer: [1-4]

1560

Now generate a new [SUBJECT] question following the same format:

Question:

1563

1564

1565

1566

1567

1568

1569

1570

Table 26: Synthetic query prompt for the News Headline Generation task.

1571

1572

Synthetic query generation prompt for News Headline Generation

1573

1574

You are a news text generator. Generate diverse news article texts that could be used to create headlines. Only output the raw news text content, not headlines or queries.

1575

1576

Example 1: [article snippet]

1577

Example 2: [article snippet]

1578

Example 3: [article snippet]

1579

Example 4: [article snippet]

1580

Example 5: [article snippet]

1581

Example 6:

1582

1583

1584

1585

1586

1587

1588

1589

1590

Table 27: Synthetic query prompt for the Scholarly Title Generation task.

1591

1592

Synthetic query generation prompt for Scholarly Title Generation

1593

1594

1595

You are a scholarly abstract generator. Generate diverse ONE paragraph abstracts that ask for creating paper titles from abstracts. The abstract should be ONE paragraph only. Only output the raw abstract text, not the actual titles.

1596

Example 1: Create a title for this research paper:

1597

Abstract: "[abstract snippet]"

1598

Title: "[title]"

1599

Example 2: Create a title for this research paper:

1600

Abstract: "[abstract snippet]"

1601

Title: "[title]"

1602

Example 3: Create a title for this research paper:

1603

Abstract: "[abstract snippet]"

1604

Title: "[title]"

1605

Example 4: Create a title for this research paper:

1606

Abstract: "[abstract snippet]"

1607

Title: "[title]"

1608

Example 5: Create a title for this research paper:

1609

Abstract: "[abstract snippet]"

1610

Title: "[title]"

1611

Example 6: Create a title for this research paper:

1612

Abstract: "[abstract snippet]"

1613

Title: "[title]"

1614

Write a research paper abstract as a single paragraph containing at least 3 sentences.

1615

1616

Example 6:

1617

1618

1619