TelcoLM: collecting data, adapting, and benchmarking language models for the telecommunication domain

Camille Barboule* **Viet-Phi Huynh*** **Adrien Bufort**
Yoan Chabot **Géraldine Damnati** **Gwénolé Lecorvé**
Orange, France
{camille.barboule,vietphi.huynh,adrien.bufort,
yoan.chabot,geraldine.damnati,gwenole.lecorve}@orange.com

Abstract

Despite outstanding processes in many tasks, Large Language Models (LLMs) still lack accuracy when dealing with highly technical domains. Especially, telecommunications (telco) is a particularly challenging domain due the large amount of lexical, semantic and conceptual peculiarities. Yet, this domain holds many valuable use cases, directly linked to industrial needs. Hence, this paper studies how LLMs can be adapted to the telco domain. It reports our effort to (i) collect a massive corpus of domain-specific data (800M tokens, 80K instructions), (ii) perform adaptation using various methodologies, and (iii) benchmark them against larger generalist models in downstream tasks that require extensive knowledge of telecommunications. Our experiments on Llama-2-7b show that domain-adapted models can challenge the large generalist models. They also suggest that adaptation can be restricted to a unique instruction-tuning step, dicarding the need for any fine-tuning on raw texts beforehand.

1 Introduction

Large Language Models like GPT-4 (openai, 2023), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023), Falcon (ZXhang, 2023) or Mistral (Jiang et al., 2023) perform very well on a wide range of tasks, on a wide range of tasks, both when the knowledge required is general and when it relates to various domains. Still, their application to highly specialized tasks, requiring advanced and up-to-date knowledge, raises a number of challenges (Zhao et al., 2023): general and frequently discussed topics tend to be disproportionately represented in their corpus (Penedo et al., 2023), whereas highly domain-specific topics tend to be underrepresented, which inevitably leads to challenges in effectively learning them for domain-specific tasks. Additionally, many domain-specific knowledge resources are proprietary assets, critical to an organization's competitive edge, and cannot be readily shared with general-purpose LLMs.

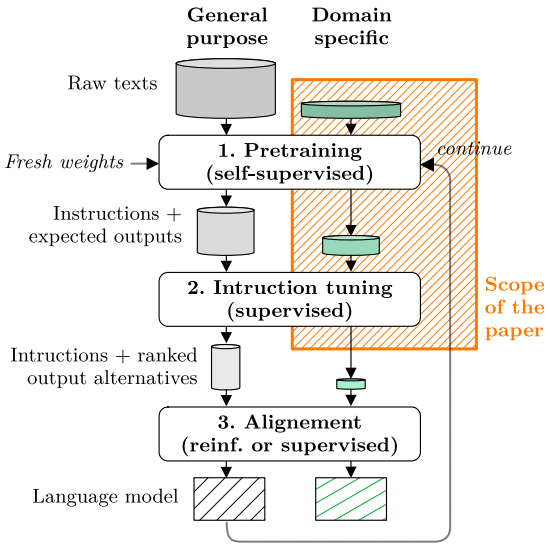

Figure 1: Steps for LM training and adaptation.

Adapting a language model to a target domain, called *domain adaptation*, is a deeply explored in the literature to solve these problems (Zhao et al., 2023). Basically, this can be performed by re-running the training steps of a baseline general-purpose LM on domain-specific data. As illustrated in Figure 1, these steps consists in (1) a pretraining step in a self-supervised manner on raw texts, (2) an instruction-tuning step on supervised datasets of instruction-output pairs, and (3) an alignement step to help the model generate texts which fits the behavioral expectations of the users.

Among domains of interests, several studies (Maatouk et al., 2023b; Bariah et al., 2023a,b) emphasize the importance of undertaking the adaptation of language models for the telecom indus-

* Main authors with equal contribution. Work achieved in March, 2024. Released in December, 2024.

try because this sector not only exhibits a large amount of lexical, semantic and concept specificities (Bariah et al., 2023b; Holm, 2021), but it is also a significant provider of natural language processing tasks, from understanding technical documents to incident resolution or network modeling (Maatouk et al., 2023b). Furthermore, in industrial applications, adapting reasonable-size models can significantly reduce computational costs, and ease deployments in environments with limited resources.

In this paper, we present experimental work to design the most efficient approach for adaptation to the telco domain. In details, our contributions concern:

- Data: We expose our collecting process to gather raw texts (800M tokens) and instructions (80k instructions) on the telco domain.
- Evaluation: We present a benchmark of telco-specific downstream tasks, combined with general-purpose evaluation to measure potential performance drops.
- Adaptation: We report a detailed comparison of various adaptation approaches by disabling or enabling some steps or some data sources.

The baseline model is Llama-2-7B (Touvron et al., 2023). The key conclusions are that the pretraining step can be skipped, and the best adapted models compete with GPT3.5.

To achieve these goals, we organize the paper as follows: Section 2 reviews the current methods' for specialized domain adaptation. Section 3 then describes the methodology used, including the data collection, how we process this domain adaptation and how we evaluate it. Experimental results including evaluation on several tasks related to telecom are then presented in Section 4. After discussing the performance of the various fine-tuning methods, we conclude the paper in Section 5.

2 Related Work

In this section, we examine the various approaches employed for addressing domain adaptation in the literature. Prominent among these are the LoRA (Hu et al., 2021), QLoRA (Dettmers et al., 2023), and ReLoRA (Lialin et al., 2023b) fine-tuning methods.

Domain adaptation without modifying the model's weights

Domain adaptation can be simply adressed using a retriever (Guu et al., 2020) which gives to the model access to a wide range of external documents. A retrieval-augmented LM allows the model to have access, in the prompt context, to an external Knowledge Base (a new corpus) (Piktus et al., 2021). This retriever enables the model to access and focus on documents within an extensive corpus, such as domain internal knowledge, giving to the model access to many sources it hasn't seen during the pretraining and fine-tuning phases. However, adding retrieval augmentation to a language model alone is insufficient for incorporating knowledge into the model. (Zhang and Choi, 2021) demonstrates that even though retrieval-augmented models were capable of updating certain knowledge when the retrieval corpus was swapped, the performance of the retrieval-augmented language model on new knowledge-related questions is notably low. In contrast, the model performs significantly better on questions related to the knowledge present in the original training corpus. The researchers also noted that substantial improvements in handling new knowledge questions only occurred after fine-tuning the model with fresh data. This indicates that merely updating the corpora from which models retrieve passages is not enough to effectively integrate knowledge into a language model. The reason behind this behavior is elucidated by the memorization process that occurs in LLMs during their pretraining phase. This memorization hinders the model's ability to effectively handle knowledge sourced from external documents via a retriever (Longpre et al., 2022). During the pretraining phase, LLMs acquire knowledge through memorization, enabling them to generate competitive results solely based on their own parametric knowledge, without the need for access to relevant documents, but (Longpre et al., 2022) demonstrates that this memorization behavior contradicts the expectation that the model should provide responses consistent with the information it retrieves, thereby reducing the system's interpretability. Most problematically, this memorization behavior severely restricts the model's ability to generalize to new knowledge not present in its training data. Furthermore, (Longpre et al., 2022) highlights that the extent of memorization during the pretraining phase increases proportionally with

Model	DAPT	Instruct. Tuning	Telco domain						General purpose			
			ATIS	3GPP	ETSI	Tele QnA	Nokia	Avg.	Open Book QA	Truthf. QA	Bigb. Narr. Und.	Avg.
Llama-2-7b			0.62	0.46	0.41	0.56	0.34	0.48	0.38	0.30	0.27	0.32
	✓		0.38	0.24	0.28	0.50	0.33	0.35	0.26	0.19	0.22	0.22
		telco	0.65	0.47	0.47	0.56	0.34	0.50	0.37	0.24	0.24	0.28
	✓	telco	0.66	0.48	0.47	0.54	**0.36**	0.50	0.36	0.28	0.22	0.29
		gen.	0.63	0.40	0.41	0.43	0.29	0.43	0.30	0.14	0.26	0.23
	✓	gen.	0.55	0.38	0.40	0.42	0.28	0.41	0.33	0.08	0.18	0.20
		gen.+telco	0.69	0.50	**0.52**	**0.61**	0.32	**0.53**	0.43	**0.33**	0.28	**0.35**
	✓	gen.+telco	**0.72**	**0.52**	0.48	0.60	0.35	**0.53**	**0.48**	0.21	**0.31**	0.33
llama-2-7b-chat			0.65	0.45	0.47	0.45	0.35	0.47	0.50	0.30	0.22	0.34
GPT3.5			0.71	0.57	0.61	0.61	0.45	0.59	0.79	0.64	0.34	0.59
GPT4			0.85	0.65	0.64	0.72	0.63	0.70	0.83	0.78	0.59	0.73

Table 6: Comparison of models accuracy on different types of MCQs. Color shades are compared to the performance of the base model (i.e., LLaMA-2 before adaptation). Red and green colors mean that the results are worse or better, respectively. Bold scores are for the best setting among the adapted LMs.

primarily because this corpus demands a broader foundational knowledge than what is provided by the technical documents and instructions utilized in the adaptation process. This discrepancy highlights a crucial aspect of domain adaptation: while it significantly enhances a model's proficiency within a specific domain, its capacity for generalization beyond the scope of the adapted materials is inherently limited.

4.2.2 MCQs results on general MCQs

We have also conducted our experiments on general datasets to analyze whether or not domain adaptation leads the model to forget general knowledge. The general MCQs benchmarks is described in 3.1.3.

Results of our experiences on these datasets can be found in Table 6. Here are our observations:

- As expected, the different adapted models do not outperform Llama-2-7b-chat-hf on General Knowledge MCQs.
- We do not observe a too-big catastrophic forgetting with our methodolody. Indeed, the adapted models have a decrease in accuracy of -0.02 with the DAPT-IAPT on general + telco instructions: it's even noteworthy that our DAPT+IAPT method, utilizing both general and telco instructions, slightly boosts performance (+0.03 in accuracy) compared to the chat model on the TruthfulQA dataset.
- Interestingly, the Llama-2-7b-IAPT and Llama-2-7b-DAPT-IAPT models, when solely based on general instructions, underperforms, possibly due to the poor quality or insufficient quantity of general instructions (40K instructions from SlimOrca (Mukherjee et al., 2023)) in the IAPT train set as the source for these general instructions).
- On text understanding MCQs, it appears our models perform competently, matching or even surpassing the baseline models. This suggests a solid grasp of broader knowledge and comprehension. This suggests that catastrophic forgetting tends to be more pronounced in certain general knowledge domains than in comprehension tasks.

4.3 Open QA results

We employ a context-free question answering (QA) approach as part of our test set to assess our models. This particular methodology, known as context-free QA, involves posing questions that do not rely on a given text or background information. For evaluating the performance of our models on this context-free QA setup, we utilize the ROUGE and METEOR metrics. ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, primarily focuses on how many of the same words and phrases appear in both the model's response and the reference answer, essentially measuring the overlap and thereby the accuracy of content. METEOR, on the other hand, stands for Metric for Evaluation of Translation with Explicit Ordering. It goes a step further by not only accounting for the similarity in terms of word overlap but also considering synonyms and the order