# Zero-Shot Cross-Lingual Summarization via Large Language Models

**Jiaan Wang**[1*†], **Yunlong Liang**[2*], **Fandong Meng**[3], **Beiqi Zou**[4]
**Zhixu Li**[5], **Jianfeng Qu**[1] **and Jie Zhou**[3]
[1]Soochow University, Suzhou, China [2]Beijing Jiaotong University, Beijing, China
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China [4]Princeton University, NJ, USA
[5]Fudan Unversity, Shanghai, China
jawang.nlp@gmail.com yunlongliang@bjtu.edu.cn
fandongmeng@tencent.com bzou@cs.princeton.edu

## Abstract

Given a document in a source language, cross-lingual summarization (CLS) aims to generate a summary in a different target language. Recently, the emergence of Large Language Models (LLMs), such as GPT-3.5, ChatGPT and GPT-4, has attracted wide attention from the computational linguistics community. However, it is not yet known the performance of LLMs on CLS. In this report, we empirically use various prompts to guide LLMs to perform zero-shot CLS from different paradigms (*i.e.*, end-to-end and pipeline), and provide a preliminary evaluation on the generated summaries. We find that ChatGPT and GPT-4 originally prefer to produce lengthy summaries with detailed information. These two LLMs can further balance informativeness and conciseness with the help of an interactive prompt, significantly improving their CLS performance. Experimental results on three widely-used CLS datasets show that GPT-4 achieves state-of-the-art zero-shot CLS performance, and performs competitively compared with the fine-tuned mBART-50.

Moreover, we also find some multi-lingual and bilingual LLMs (*i.e.*, BLOOMZ, ChatGLM-6B, Vicuna-13B and ChatYuan) have limited zero-shot CLS ability. Due to the composite nature of CLS, which requires models to perform summarization and translation simultaneously, accomplishing this task in a zero-shot manner is even a challenge for LLMs. *Therefore, we sincerely hope and recommend future LLM research could use CLS as a testbed.*

## 1 Introduction

Cross-Lingual Summarization (CLS) aims to provide a target-language (*e.g.*, Chinese) summary for a lengthy document in a different source language (*e.g.*, English) (Leuski et al., 2003; Wan et al., 2010; Yao et al., 2015; Zhu et al., 2019, 2020; Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021; Bai et al., 2021a; Liang et al., 2022b; Feng et al., 2022; Hasan et al., 2021; Wang et al., 2022a,b; Liang et al., 2022a; Liu et al., 2022; Zheng et al., 2022; Aumiller et al., 2022). This task could help people quickly capture their interests from foreign documents.

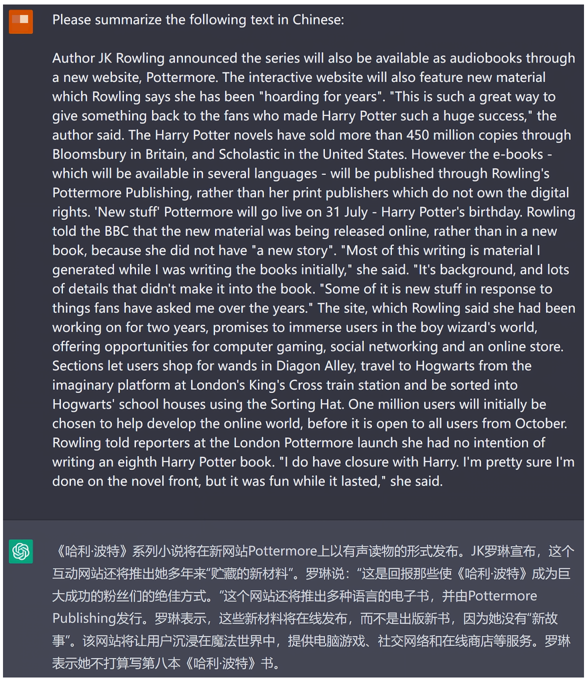



Figure 1: An example of zero-shot CLS via ChatGPT.

In recent years, a number of powerful multi-lingual pre-trained generative models have been proposed one after another, such as mBART (Liu et al., 2020), mBART-50 (Tang et al., 2021), mT5 (Xue et al., 2021) and BLOOM (Scao et al., 2022). The parameters in these models have gradually increased from million levels (*e.g.*, 580M in mT5-base and 610M in mBART-Large) to billion levels (*e.g.*, 3.7B in mT5-XL, 13B in mT5-XXL and 176B in BLOOM), facilitating various research topics (*e.g.*, machine translation and CLS) in the multi-lingual world. Besides, large language mod-

[*] Equal Contribution. Work was done when Wang and Liang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
[†] Corresponding author.

els (LLMs) have been key to strong performance when transferring to new tasks by simply conditioning on a few input-label pairs (*in-context learning*) (Dong et al., 2022; Min et al., 2022) or short sentences describing crucial reasoning steps (*chain-of-thoughts*) (Fu et al., 2022; Zhang et al., 2022).

More recently, ChatGPT and GPT-4 (OpenAI, 2023) have attracted great attention from both the research communities and industries. Similar to InstructGPT (Ouyang et al., 2022), ChatGPT is created by fine-tuning a GPT-3.5 series model via reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). GPT-4, as a multi-modal LLM that can accept image and text inputs and produce text outputs, exhibits human-level performance on various benchmark datasets (OpenAI, 2023). With the emergence of ChatGPT and GPT-4, there is growing interest in leveraging LLMs for various NLP tasks (Qin et al., 2023; Jiao et al., 2023; Bang et al., 2023; Yang et al., 2023; Zhong et al., 2023; Wang et al., 2023; Bubeck et al., 2023; Tan et al., 2023; Peng et al., 2023; Liu et al., 2023; Yong et al., 2023). Nevertheless, the exploration of LLMs on CLS is still lacking.

In this report, we present a preliminary evaluation of LLMs' zero-shot CLS performance, including GPT-3.5, ChatGPT, GPT-4, BLOOMZ, ChatGLM-6B, Vicuna-13B and ChatYuan. In detail, we design various prompts to guide LLMs to perform CLS in an end-to-end manner with or without chain-of-thoughts (CoT). Figure 1 gives an example of prompting ChatGPT to perform zero-shot CLS. To further exploit the interaction capability of conversational LLMs (*e.g.*, ChatGPT and GPT-4), we leverage an interactive prompt to let them produce more concise summaries. Moreover, to provide a deeper analysis of LLMs' zero-shot CLS performance, we compare them with fine-tuned mBART-50 (Tang et al., 2021) which has shown its superiority in many previous CLS works (Wang et al., 2022a; Feng et al., 2022; Perez-Beltrachini and Lapata, 2021). Experimental results on three CLS datasets, covering three domains (news, how-to guide and dialogue) and two cross-lingual directions (En⇒Zh and En⇒De)[1], show that GPT-4 achieves the best zero-shot performance but is still worse than the fine-tuned mBART-50 model in terms of ROUGE scores and BERTScore. We also conduct case studies to show that ChatGPT and GPT-4 could absorb the core idea of the given source-language documents and generate fluent and concise target-language summaries.

In addition, we find that the current open-source LLMs (*i.e.*, BLOOMZ, ChatGLM-6B, Vicuna-13B and ChatYuan) achieve limited zero-shot CLS performance, which is significantly worse than that of GPT-4. This leads us to conclude that the composite end-to-end CLS prompts are difficult for them to follow, and there is still a challenge for LLMs to perform zero-shot CLS in an end-to-end manner which requires simultaneously carrying out translation and summarization. Based on the finding, we suggest that future multi-lingual or bilingual LLM research uses CLS as a testbed to evaluate LLMs' capabilities to follow composite instructions as well as combine their different abilities.

Our main contributions are concluded as follows:

- To the best of our knowledge, we are the first to explore the zero-shot CLS performance of LLMs. To achieve that, we design various prompts to guide LLMs to perform CLS in an end-to-end manner with or without chain-of-thoughts.
- Experimental results on three widely-used CLS benchmark datasets, covering various domains and languages, show several LLMs (especially ChatGPT and GPT-4) achieve competitive results compared with the strong fine-tuned baseline.
- We also find the current open-source LLMs generally achieve limited zero-shot CLS performance, making us think CLS could be used as a testbed for future LLM research due to its challenges.

## 2 Methodology

### 2.1 Cross-Lingual Summarization Prompts

We heuristically design 3 prompts to guide LLMs to perform zero-shot CLS in an end-to-end manner, which is shown as follows with an example from an English document to a Chinese summary:

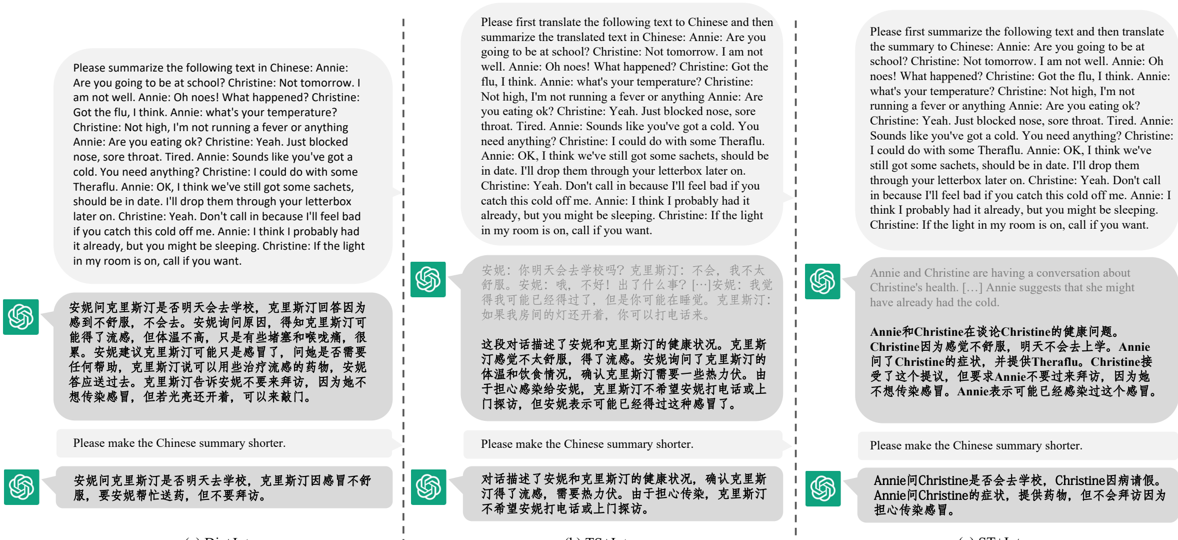
- The direct (**Dir**) prompt guides LLMs straightforwardly output the corresponding target-language summary without chain-of-thoughts (CoT):

```
Please summarize the
following text in
Chinese:  [English Doc]
```

  where `[English Doc]` indicates a given English document.

- The translate-then-summarize (**TS**) CoT prompt

[1] Since a CLS dataset might contain multiple source and target languages, we use "X⇒Y" to indicate the source language and target language are X and Y, respectively. En: English; Zh: Chinese; De: German.




(a) Dir+Inta (b) TS+Inta (c) ST+Inta

Figure 2: An illustration of all prompts used to guide LLMs to perform zero CLS (take ChatGPT as an example).

makes LLMs first translate the given document from the source language to the target language, and then summarize the translated document to perform CLS:

```
Please first translate
the following text
to Chinese and then
summarize the translated
text in Chinese:
[English Doc]
```

- The summarize-then-translate (**ST**) CoT prompt lets LLMs first summarize the given document and then translate the output summary to the target language:

```
Please first summarize
the following text
and then translate the
summary to Chinese:
[English Doc]
```

Note that though the TS and ST CoT prompts guide LLMs to perform CLS step by step, the behaviors are end-to-end since the target-language summaries are generated within a single turn.

To further exploit the potentiality of conversational LLMs, inspired by Bang et al. (2023), after prompting with Dir, TS or ST prompt, we adopt an interactive (**Inta**) prompt to make the preliminarily generated summary more concise:

```
Please make the Chinese
summary shorter.
```

and the whole process is denoted as "**Dir+Inta**", "**TS+Inta**" or "**ST+Inta**".

### 2.2 Large Language Models

We explore the CLS ability of the following LLMs:

- **Davinci-003** is the most advanced GPT-3.5 model with 175B parameters. We evaluate its performance by requesting the official API provided by OpenAI with default settings.[2]
- **ChatGPT** is created by fine-tuning a GPT-3.5 series model via reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). We conduct experiments on the ChatGPT platform[3] between February 17 to February 19, 2023.
- **GPT-4**, as a multi-modal LLM that can accept image and text inputs and produce text outputs, exhibits human-level performance on various benchmark datasets (OpenAI, 2023). We assess GPT-4 on the ChatGPT platform between March 15 to March 19, 2023.
- **BLOOMZ** (Muennighoff et al., 2022) is an open-source multi-lingual LLM with 176B parameters. The model supports 59 languages, and is created by fine-tuning BLOOM (Scao et al., 2022) on an instruction corpus (*i.e.*, XP3).
- **ChatGLM-6B**[4] is an open-source bilingual (*i.e.*, Chinese and English) language model based on General Language Model (GLM) framework (Du

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://chat.openai.com/

[4] https://github.com/THUDM/ChatGLM-6B

| Dataset | Src Lang. | Trg Lang. | Domain | Example | Doc. Length | Sum. Length |
|---|---|---|---|---|---|---|
| CrossSum | English | Chinese | News | 3981 / 497 / 50 out of 497 | 814.2 | 35.6 |
| WikiLingua | English | Chinese | How-to guide | 13211 / 1886 / 50 out of 3775 | 538.6 | 53.2 |
| | | German | | 40839 / 5833 / 50 out of 11669 | 526.1 | 63.4 |
| XSAMSum | English | Chinese | Dialogue | 14732 / 818 / 50 out of 819 | 140.1 | 27.6 |
| | | German | | 14732 / 818 / 50 out of 819 | 140.1 | 31.7 |

Table 1: Statistics of CLS datasets used in experiments. "*Src Lang.*" and "*Trg Lang*" denote the source and the target languages. "*Doc. Length*" and "*Sum. Length*" show the average length of source documents and target summaries (token level). "*Example*" lists the number of samples in each dataset w.r.t training, validation and test sets.

et al., 2022). The model suffers from both instruction tuning and RLHF.

- **Vicuna-13B**[5] is an open-source LLM created by fine-tuning LLaMA (Touvron et al., 2023) on user-shared conversations collected from ChatGPT. We evaluate the model via its demo platform[6] between March 31 to April 2, 2023.
- **ChatYuan**[7] is an open-source bilingual (*i.e.*, Chinese and English) LLM with 7.7B parameters. The training process of this model includes instruction tuning and RLHF.

Among the above LLMs, ChatGPT, GPT-4, ChatGLM-6B, Vicuna-13B and ChatYuan are conversational LLMs while Davinci-003 and BLOOMZ are not. When evaluating their zero-shot CLS performance, we only equip conversational LLMs with the interactive prompt.

# 3 Experiments

## 3.1 Experimental Setup

**Datasets.** We evaluate LLMs on the following three CLS datasets: CrossSum (En⇒Zh) (Hasan et al., 2021), WikiLingua (En⇒Zh/De) (Ladhak et al., 2020) and XSAMSum (En⇒Zh/De) (Wang et al., 2022a). CrossSum is collected from BBC news website, it contains 3,981 English news reports paired with Chinese summaries. WikiLingua involves 18,887 English how-to guides paired with Chinese summaries, and 58,375 English how-to guides paired with German summaries. Note that both CrossSum and WikiLingua also provide CLS samples in other cross-lingual directions, and we only utilize En⇒Zh or (and) En⇒De samples in this work. XSAMSum contains 16,369 English dialogues paired with both Chinese and German summaries. The detailed statistics of these datasets are listed in Table 1. Since ChatGPT, GPT-4 and Vicuna-13B can only be interacted with manually when we conduct experiments, evaluating their performance is time-consuming. Thus, we randomly sample 50 documents from the test set of each CLS dataset for evaluation.

**Metrics.** We adopt ROUGE-1/2/L (R-1/2/L) (Lin, 2004) and BERTScore (B-S) (Zhang et al., 2020) in our experiments. The ROUGE scores measure the lexical overlap between the generated summaries and corresponding references based on the unigram, bigram and longest common subsequence, while the BERTScore measures the semantic similarity. For ROUGE scores, we use *multi-lingual rouge*[8] toolkit. For BERTScore, we use *bert-score*[9] toolkit, and the score is calculated based on *bert-base-multilingual-cased*[10] model.

**Baselines.** We also compare zero-shot LLMs with fine-tuned mBART-50 (Tang et al., 2021) to provide a deeper analysis. mBART-50 is a multi-lingual version of BART (Lewis et al., 2020) with the vanilla transformer encoder-decoder architecture (Vaswani et al., 2017). This model has been pre-trained on large-scale multi-lingual unlabeled corpora with BART-like denoising objectives.

## 3.2 Implementation Details

For ChatGPT, GPT-4 and Vicuna-13B, we manually evaluate their results via the corresponding platform and demo websites. Among them, the demo website of Vicuna-13B cannot support the long input sequences, and it will automatically truncate the long sequences, thus we only evaluate Vicuna-13B on XSAMSum (En⇒Zh/De). For Davinci-003, we use the official API with default settings.

[5] https://vicuna.lmsys.org/
[6] https://chat.lmsys.org/
[7] https://github.com/clue-ai/ChatYuan
[8] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring
[9] https://github.com/Tiiiger/bert_score
[10] https://huggingface.co/bert-base-multilingual-cased

| Model | CrossSum (En⇒Zh) | | | | WikiLingua (En⇒Zh) | | | | WikiLingua (En⇒De) | | | | XSAMSum (En⇒Zh) | | | | XSAMSum (En⇒De) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | B-S | R-1 | R-2 | R-L | B-S | R-1 | R-2 | R-L | B-S | R-1 | R-2 | R-L | B-S | R-1 | R-2 | R-L | B-S |
| mBART-50 | 26.1 | 7.4 | 22.1 | 65.4 | 32.1 | 10.4 | 26.8 | 68.5 | 26.8 | 7.7 | 20.5 | 62.5 | 40.6 | 14.4 | 33.9 | 74.5 | 42.4 | 18.9 | 35.4 | 73.7 |
| ChatYuan-7.7B (Dir) | 0.3 | 0.0 | 0.3 | 51.7 | 4.1 | 1.2 | 2.6 | 54.4 | - | - | - | - | 0.8 | 0.3 | 0.7 | 48.5 | - | - | - | - |
| ChatYuan-7.7B (Dir+Inta) | 0.2 | 0.0 | 0.2 | 52.0 | 4.7 | 1.6 | 3.4 | 51.6 | - | - | - | - | 0.3 | 0.1 | 0.3 | 47.0 | - | - | - | - |
| ChatYuan-7.7B (TS) | 0.4 | 0.0 | 0.4 | 46.6 | 8.2 | 2.7 | 5.4 | 56.0 | - | - | - | - | 11.3 | 4.2 | 8.7 | 49.5 | - | - | - | - |
| ChatYuan-7.7B (TS+Inta) | 2.0 | 0.5 | 1.4 | 46.9 | 6.9 | 2.1 | 4.3 | 53.4 | - | - | - | - | 9.5 | 3.2 | 6.9 | 52.3 | - | - | - | - |
| ChatYuan-7.7B (ST) | 0.5 | 0.0 | 0.4 | 49.6 | 6.9 | 2.1 | 4.2 | 56.1 | - | - | - | - | 7.5 | 2.5 | 5.5 | 49.6 | - | - | - | - |
| ChatYuan-7.7B (ST+Inta) | 1.2 | 0.4 | 0.9 | 49.7 | 7.3 | 2.3 | 4.6 | 55.5 | - | - | - | - | 6.0 | 2.0 | 3.9 | 48.5 | - | - | - | - |
| ChatGLM-6B (Dir) | 5.7 | 2.3 | 2.4 | 53.9 | 14.5 | 5.3 | 9.9 | 59.5 | - | - | - | - | 20.4 | 9.1 | 15.3 | 58.8 | - | - | - | - |
| ChatGLM-6B (Dir+Inta) | 7.9 | 2.4 | 5.3 | 55.6 | 14.6 | 5.1 | 9.5 | 59.1 | - | - | - | - | 18.0 | 8.0 | 14.0 | 59.5 | - | - | - | - |
| ChatGLM-6B (TS) | 8.4 | 2.9 | 4.8 | 54.1 | 14.6 | 5.3 | 9.8 | 59.7 | - | - | - | - | 21.5 | 9.6 | 16.6 | 57.9 | - | - | - | - |
| ChatGLM-6B (TS+Inta) | 9.6 | 3.0 | 6.1 | 55.2 | 14.9 | 5.1 | 9.4 | 59.1 | - | - | - | - | 18.7 | 8.1 | 15.0 | 58.6 | - | - | - | - |
| ChatGLM-6B (ST) | 5.8 | 1.8 | 3.6 | 53.2 | 15.6 | 5.5 | 10.2 | 59.9 | - | - | - | - | 19.8 | 8.3 | 14.7 | 58.1 | - | - | - | - |
| ChatGLM-6B (ST+Inta) | 2.2 | 0.6 | 1.7 | 53.8 | 9.8 | 3.3 | 6.1 | 57.0 | - | - | - | - | 12.7 | 5.1 | 9.9 | 56.8 | - | - | - | - |
| Vicuna-13B (Dir) | - | - | - | - | - | - | - | - | - | - | - | - | 19.5 | 7.2 | 14.5 | 60.1 | 22.5 | 4.9 | 17.6 | 58.5 |
| Vicuna-13B (Dir+Inta) | - | - | - | - | - | - | - | - | - | - | - | - | 24.1 | 9.7 | 18.9 | 63.0 | 28.7 | 7.8 | 22.0 | 60.5 |
| Vicuna-13B (TS) | - | - | - | - | - | - | - | - | - | - | - | - | 18.3 | 7.1 | 14.6 | 61.6 | 25.0 | 5.9 | 18.2 | 59.4 |
| Vicuna-13B (TS+Inta) | - | - | - | - | - | - | - | - | - | - | - | - | 22.0 | 7.9 | 17.4 | 64.3 | 31.7 | 8.9 | 24.2 | 61.2 |
| Vicuna-13B (ST) | - | - | - | - | - | - | - | - | - | - | - | - | 17.5 | 6.1 | 13.6 | 59.6 | 27.3 | 6.8 | 20.4 | 59.3 |
| Vicuna-13B (ST+Inta) | - | - | - | - | - | - | - | - | - | - | - | - | 19.8 | 7.4 | 15.4 | 62.2 | 31.6 | 9.4 | 24.1 | 61.9 |
| BLOOMZ-176B (Dir) | 0.7 | 0.1 | 0.7 | 29.2 | 0.3 | 0.0 | 0.2 | 8.9 | 0.0 | 0.0 | 0.0 | 3.3 | 21.4 | 11.2 | 17.8 | 65.3 | 13.0 | 1.2 | 11.9 | 56.2 |
| BLOOMZ-176B (TS) | 2.1 | 1.3 | 1.6 | 21.5 | 0.4 | 0.0 | 0.3 | 5.6 | 0.0 | 0.0 | 0.0 | 5.0 | 30.4 | 15.0 | 25.2 | 64.8 | 12.5 | 0.7 | 11.4 | 54.4 |
| BLOOMZ-176B (ST) | 3.0 | 1.2 | 2.4 | 33.8 | 0.3 | 0.0 | 0.2 | 9.0 | 0.0 | 0.0 | 0.0 | 3.3 | 28.1 | 13.4 | 23.4 | 66.3 | 13.8 | 1.3 | 12.8 | 54.8 |
| Davinci-003 (Dir) | 18.7 | 3.6 | 14.7 | 60.2 | 23.6 | 3.8 | 17.8 | 60.9 | 18.8 | 2.6 | 12.2 | 60.7 | 24.4 | 8.0 | 20.7 | 63.4 | 35.5 | 12.4 | 27.3 | 62.4 |
| Davinci-003 (TS) | 22.9 | 8.9 | 13.5 | 59.6 | 23.7 | 8.2 | 15.1 | 61.0 | 16.9 | 2.0 | 10.9 | 59.2 | 33.3 | 17.1 | 26.6 | 64.7 | 34.7 | 11.5 | 26.1 | 62.0 |
| Davinci-003 (ST) | **26.2** | **9.3** | **16.9** | 61.3 | 24.2 | 8.4 | 15.9 | 61.2 | 19.8 | 2.8 | 13.1 | 60.4 | 34.1 | **18.2** | 26.4 | 68.1 | 35.7 | 11.7 | 26.9 | 63.0 |
| ChatGPT (Dir) | 14.2 | 3.3 | 10.3 | 60.3 | 20.9 | 5.6 | 15.5 | 62.7 | 16.9 | 2.1 | 10.7 | 60.1 | 21.3 | 5.5 | 17.1 | 63.5 | 32.0 | 10.3 | 24.5 | 61.4 |
| ChatGPT (Dir+Inta) | 22.1 | 3.8 | 15.6 | 61.8 | 28.4 | 6.5 | 22.1 | 64.5 | 22.4 | 2.8 | 14.7 | 61.3 | 27.2 | 6.9 | 22.9 | 67.5 | 39.6 | **16.0** | 31.4 | **64.3** |
| ChatGPT (TS) | 15.8 | 3.3 | 11.9 | 60.9 | 24.8 | 5.4 | 19.1 | 62.9 | 19.4 | 2.4 | 12.6 | 60.0 | 26.0 | 7.3 | 21.2 | 66.4 | 33.2 | 9.6 | 25.3 | 61.1 |
| ChatGPT (TS+Inta) | 22.6 | 4.1 | **16.9** | 62.7 | 26.1 | 5.3 | 19.7 | 63.7 | 21.6 | 2.4 | 15.1 | 60.8 | 27.4 | 6.7 | 22.4 | 67.1 | 39.4 | 13.5 | 29.4 | 63.3 |
| ChatGPT (ST) | 16.5 | 3.8 | 12.0 | 60.8 | 27.2 | 7.3 | 20.3 | 64.3 | 21.3 | 3.5 | 14.4 | 60.9 | 26.8 | 7.7 | 21.3 | 66.7 | 31.7 | 8.8 | 23.5 | 60.8 |
| ChatGPT (ST+Inta) | 21.6 | 3.5 | 15.5 | 61.7 | 30.1 | 8.1 | **22.4** | 64.9 | 21.4 | 3.1 | 15.4 | 60.6 | 31.4 | 11.5 | 28.1 | 70.1 | 35.9 | 13.2 | 29.0 | 62.8 |
| GPT-4 (Dir) | 13.7 | 3.7 | 10.1 | 59.7 | 23.1 | 9.1 | 15.5 | 63.5 | 20.4 | 3.3 | 13.8 | 62.2 | 24.5 | 7.1 | 19.5 | 66.1 | 34.7 | 13.4 | 25.3 | 61.7 |
| GPT-4 (Dir+Inta) | 20.3 | 4.4 | 14.1 | 61.9 | 30.4 | 11.7 | 20.9 | **65.7** | **24.8** | 3.9 | 17.0 | **63.5** | 31.3 | 7.3 | 26.5 | 70.7 | **40.5** | 13.4 | 30.8 | 64.2 |
| GPT-4 (TS) | 19.4 | 3.6 | 14.3 | 60.9 | 28.5 | 11.4 | 18.2 | 64.2 | 23.1 | 3.8 | 16.3 | 62.7 | **34.7** | 12.5 | **28.5** | 71.0 | 38.9 | 11.9 | 29.0 | 63.3 |
| GPT-4 (TS+Inta) | 22.7 | 4.3 | 16.1 | 62.2 | 29.2 | **12.6** | 20.3 | 64.9 | 23.6 | 3.9 | 17.3 | 62.9 | 30.8 | 6.5 | 25.6 | 70.9 | 39.1 | 13.5 | **32.6** | 64.1 |
| GPT-4 (ST) | 19.0 | 4.3 | 14.1 | 61.7 | 30.2 | 12.2 | 19.5 | 64.2 | 23.4 | 3.8 | 16.4 | 63.0 | 32.1 | 10.7 | 26.4 | 70.7 | 38.6 | 12.3 | 29.5 | 63.2 |
| GPT-4 (ST+Inta) | 22.6 | 4.9 | 16.8 | **63.1** | **30.5** | 11.9 | 21.3 | 65.2 | 23.1 | **4.2** | **17.4** | 62.7 | 29.2 | 8.2 | 25.4 | **71.4** | 39.0 | 11.5 | 31.2 | 63.7 |

Table 2: Experimental results on CrossSum, WikiLingua and XSAMSum. Pink denotes the fine-tuned baseline. Light blue and blue denote the zero-shot performance of open-source and non-open-source LLMs, respectively. Green indicates the zero-shot result is better than that of the fine-tuned baseline. "-" denotes the model cannot be evaluated in the corresponding dataset.

For BLOOMZ, ChatGLM-6B and ChatYuan, we download the corresponding checkpoints and evaluate their performances following the officially released codes. The 176B BLOOMZ makes use of 5*80G GPUs to load with FP16 precision. We use a sampling decoding strategy and set the temperature to 0.7. We only evaluate ChatGLM-6B and ChatYuan on En⇒Zh cross-lingual direction due to their bilingualism (*i.e.*, Chinese and English).

For mBART-50 baseline, inspired by Feng et al. (2022) and Wang et al. (2022a), we employ `mBART-large-50-many-to-many-mmt` model[11] via the implementation of Huggingface Transformers (Wolf et al., 2020). This model is fine-tuned with 4 batch size, 5e-6 learning rate and 10 epochs on each CLS dataset.

[11] `https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt`

### 3.3 Main Results

Table 2 lists the experimental results. As we can see, Davinci-003, ChatGPT and GPT-4 achieve competitive results with the fine-tuned mBART-50.

**The Effect of Each CLS Prompt.** Among three end-to-end prompts (*i.e.*, Dir, ST and TS), the CoT prompts lead to better performance than the direct prompt, indicating the effectiveness of CoT. It also indicates that it is still challenging for a single model to directly perform CLS without giving any crucial or helpful instructions.

**The Effect of Interactive Prompt.** Further, with the help of the interactive prompt, the performance of ChatGPT and GPT-4 significantly improve and even outperform mBART-50 in several automatic metrics. As shown in Table 3, more concise summaries can be generated after inputting the interactive prompt, *e.g.*, 183.7 tokens generated by ChatGPT (Dir) on CrossSum, while the counterpart of ChatGPT (Dir+Inta) is 66.4 tokens. Figure 3

| English Dialogue | Generated Summaries |
|---|---|
| Boris: Due to the number of PTO requests this week, we had to decide who will be allowed to take off based on seniority.<br>Sarah: I understand.<br>Boris: Unfortunately, you are too low on the totem pole and we will have to deny your request.<br>Sarah: Oh, that's too bad! Is there any compromise? I really wanted to have some time with my family.<br>Boris: I can give you first choice of dates the week after, but that's the best I can do.<br>Sarah: What if someone cancels?<br>Boris: That is a rare, but yes, I will come to you first.<br>Sarah: Thank you.<br>Boris: Again, I'm very sorry.<br>Sarah: Could you amend my request for the following week, Monday through Wednesday?<br>Boris: Certainly. That should be no problem.<br>Sarah: That's great, at least I'll have a long weekend!<br>Boris: Very true. No problems from my side.<br>Sarah: Do I need to initial the change?<br>Boris: Yes, please come see my assistant at your earliest convenience.<br>Sarah: Certainly, no problem.<br>Boris: Thank you for being so flexible.<br>Sarah: Sure, I understand we can't all be off. You are just trying to be fair.<br>Boris: That's it exactly. Sarah: Thank you again for explaining it so well.<br>Boris: My pleasure. Have a nice holiday and enjoy your time off. Sarah: Thank you! | **ChatGPT (Dir):** Boris通知Sarah由于请假请求众多，公司不得不按资历高低来决定谁可以休假。由于Sarah资历不够高，公司不得不拒绝她的请求。Sarah想知道是否有妥协的余地，Boris答应给她下周第一批可以选择的日期，但这是他所能做的最好安排。Sarah问如果有人取消假期怎么办，Boris说这是很少见的，但如果确实发生，他会优先通知Sarah。最后，Sarah改变了请假日期，Boris同意了，并建议她尽快与他的助手联系确认。他们互相道别，Sarah感谢Boris的解释，Boris祝她假期愉快。<br>**ChatGPT (Dir+Inta):** Sarah想请假，但因资历不够高被拒绝。Boris答应给她下周第一批可以选择的日期，如果有人取消假期，会优先通知她。最后，Sarah改变了请假日期，Boris同意了。<br>**ChatGPT (TS):** 对于申请休假的员工，由于请假申请数量较多，公司根据资历来决定谁可以休假，导致低资历的员工被拒绝休假请求。员工虽然表示遗憾，但同意公司的决定并愿意接受推迟的休假时间，直到后来有机会更改休假时间并获得了首选日期。公司对员工的灵活性表示感谢，并向员工祝福愉快的假期。<br>**ChatGPT (TS+Inta):** 公司根据资历决定员工休假，低资历员工被拒绝休假请求。员工同意公司决定并愿意推迟休假时间，后来获得了首选日期。公司感谢员工的灵活性，祝愿员工假期愉快。<br>**ChatGPT (ST):** 鲍里斯告诉萨拉，由于请假申请太多，他必须根据资历批准休假，但很遗憾，萨拉的请求不能被批准。然而，鲍里斯提供萨拉选择下周的日期，如果有人取消。萨拉问能否修改为下周的请求，鲍里斯同意了。他们以积极的态度结束了谈话，鲍里斯祝萨拉度过愉快的假期，萨拉感谢他的理解。<br>**ChatGPT (ST+Inta):** 因请假申请太多，鲍里斯按资历批准休假，未能批准萨拉的请求。但鲍里斯提供了选择下周日期的机会。萨拉同意改为下周休假，他们友好结束谈话。<br>**GPT-4 (Dir):** Boris通知Sarah由于请假请求众多，公司不得不按资历高低来决定谁可以休假。由于Sarah资历不够高，公司不得不拒绝她的请求。Sarah想知道是否有妥协的余地，Boris答应给她下周第一批可以选择的日期，但这是他所能做的最好安排。Sarah问如果有人取消假期怎么办，Boris说这是很少见的，但如果确实发生，他会优先通知Sarah。最后，Sarah改变了请假日期，Boris同意了，并建议她尽快与他的助手联系确认。他们互相道别，Sarah感谢Boris的解释，Boris祝她假期愉快<br>**GPT-4 (Dir+Inta):** Sarah想请假，但因资历不够高被拒绝。Boris答应给她下周第一批可以选择的日期，如果有人取消假期，会优先通知她。最后，Sarah改变了请假日期，Boris同意了。<br>**GPT-4 (TS):** 由于本周请假申请众多，鲍里斯根据员工资历决定谁能休假，萨拉的申请被拒。经过沟通，鲍里斯同意在下周给萨拉优先选择日期，若有人取消，也会优先考虑萨拉。最后，鲍里斯帮助萨拉将申请更改为下周一至周三。萨拉表示理解公平性并感谢鲍里斯的帮助。<br>**GPT-4 (TS+Inta):** 鲍里斯因员工请假申请多而根据资历拒绝了萨拉的请求。经沟通，鲍里斯同意下周给萨拉优先选择日期，若有人取消，也会优先考虑。萨拉表示理解公平性并感谢鲍里斯。<br>**GPT-4 (ST):** 由于本周休假申请数量众多，鲍里斯不得不根据资历来决定谁可以休假，因此拒绝了萨拉的申请。他们讨论了潜在的妥协方案，最终同意萨拉在下周一至周三休假。如果有人取消休假，鲍里斯将首先向萨拉提供名额。萨拉需要在鲍里斯的助手那里签署更改。<br>**GPT-4 (ST+Inta):** 鲍里斯因资历问题拒绝了萨拉的休假申请。他们商量后同意萨拉下周一至三休假。若有人取消，鲍里斯会优先通知萨拉。萨拉需在助手处确认更改。 |

Figure 3: Example dialogue document in XSAMSum and summaries generated by ChatGPT and GPT-4.

also shows an example English document with the corresponding summaries generated by ChatGPT and GPT-4 via different prompts. Therefore, the conversational LLMs prefer to generate lengthy summaries probably due to the RLHF training process, and the interactive prompt further helps them balance informativeness and conciseness, and significantly improves their zero-shot CLS ability.

**Best Zero-Shot LLM vs. Fine-Tuned mBART.** GPT-4 achieves state-of-the-art zero-shot CLS performance among all LLMs, justifying its superiority. But the model is still slightly worse than the fine-tuned mBART-50 in terms of automatic evaluation metrics. One possible reason is that zero-shot LLMs are not aware of the text style of the golden summaries when performing zero-shot CLS on each dataset. However, lower automatic scores do not indicate worse performance. For example, as discussed by Goyal et al. (2022), the news summaries generated by GPT-3 achieve lower ROUGE scores than fine-tuned models but higher in human evaluation. Thus, the comparison between LLMs and fine-tuned mBART-50 in CLS needs human evaluation, which we reserve for the future.

**Limited Performance of Open-Source LLMs.** For open-source LLMs, *i.e.*, BLOOMZ, ChatGLM-6B, Vicuna-13B and ChatYuan-7.7B, they perform poorly on CrossSum and WikiLingua datasets whose documents typically contain more lengthy content than those of XSAMSum. Although they perform decently on XSAMSum, there is still a large gap compared to GPT-4. Thus, we conclude that zero-shot CLS is challenging for LLMs to perform due to its composite nature that requires models to perform summarization and translation simultaneously. In this situation, we suggest future bilingual or multi-lingual LLM research adopt CLS as a testbed to evaluate the LLMs' capabilities to follow composite instructions as well as combine their different ability.

| Method | CrossSum | WikiLingua | | XSAMSum | |
|---|---|---|---|---|---|
| | En⇒Zh | En⇒Zh | En⇒De | En⇒Zh | En⇒De |
| mBART-50 | 32.7 | 46.6 | 75.4 | 22.3 | 27.9 |
| Davinci-003 (Dir) | 83.3 | 78.5 | 149.1 | 61.8 | 62.5 |
| Davinci-003 (TS) | 82.1 | 76.2 | 148.6 | 53.4 | 65.8 |
| Davinci-003 (ST) | 44.7 | 49.1 | 91.7 | 43.4 | 52.1 |
| ChatGPT (Dir) | 183.7 | 176.6 | 273.5 | 68.6 | 75.3 |
| ChatGPT (Dir+Inta) | 66.4 | 50.0 | 80.7 | 28.7 | 42.5 |
| ChatGPT (TS) | 155.1 | 82.1 | 149.3 | 48.2 | 60.9 |
| ChatGPT (TS+Inta) | 63.4 | 46.2 | 70.0 | 30.3 | 41.1 |
| ChatGPT (ST) | 132.7 | 94.3 | 124.2 | 54.9 | 68.1 |
| ChatGPT (ST+Inta) | 57.8 | 50.1 | 71.6 | 29.3 | 37.5 |
| GPT-4 (Dir) | 227.1 | 170.5 | 193.1 | 70.4 | 74.4 |
| GPT-4 (Dir+Inta) | 102.2 | 58.7 | 75.1 | 30.1 | 38.3 |
| GPT-4 (TS) | 93.9 | 85.6 | 114.7 | 44.1 | 53.8 |
| GPT-4 (TS+Inta) | 56.5 | 45.4 | 66.5 | 26.3 | 33.8 |
| GPT-4 (ST) | 106.6 | 87.8 | 109.5 | 43.6 | 53.7 |
| GPT-4 (ST+Inta) | 62.7 | 48.0 | 65.1 | 26.7 | 33.3 |
| **Golden** | 36.1 | 50.0 | 66.8 | 23.9 | 29.6 |

Table 3: The average length (token level) of the generated summaries on the test set of each CLS dataset. Light green indicates the length of golden summaries.

### 3.4 LLM-based Evaluation

It is worth noting that conducting human evaluation on the generated summaries of both LLMs and fine-tuned models is not trivial since human evaluators

| | XSAMSum (En⇒Zh) | | | | XSAMSum (En⇒De) | | | |
|---|---|---|---|---|---|---|---|---|
| | Coherence | Relevance | Consistency | Fluency | Coherence | Relevance | Consistency | Fluency |
| mBART-50 | 54.0 | 32.3 | 36.6 | 55.8 | 54.6 | 36.0 | 45.3 | 52.1 |
| ChatYuan-7.7B (Dir) | 44.6 | 17.3 | 40.3 | 53.6 | - | - | - | - |
| ChatYuan-7.7B (Dir+Inta) | 43.6 | 21.0 | 37.3 | 52.0 | - | - | - | - |
| ChatYuan-7.7B (TS) | 41.0 | 17.0 | 24.8 | 48.8 | - | - | - | - |
| ChatYuan-7.7B (TS+Inta) | 39.0 | 15.3 | 17.3 | 41.5 | - | - | - | - |
| ChatYuan-7.7B (ST) | 47.3 | 19.6 | 36.3 | 54.6 | - | - | - | - |
| ChatYuan-7.7B (ST+Inta) | 48.6 | 15.3 | 33.0 | 52.0 | - | - | - | - |
| ChatGLM-6B (Dir) | 58.8 | 31.0 | 49.0 | 61.0 | - | - | - | - |
| ChatGLM-6B (Dir+Inta) | 60.6 | 35.3 | 55.1 | 60.8 | - | - | - | - |
| ChatGLM-6B (TS) | 52.0 | 22.0 | 25.3 | 54.0 | - | - | - | - |
| ChatGLM-6B (TS+Inta) | 55.0 | 31.6 | 46.5 | 58.5 | - | - | - | - |
| ChatGLM-6B (ST) | 58.6 | 27.0 | 37.3 | 56.5 | - | - | - | - |
| ChatGLM-6B (ST+Inta) | 59.3 | 34.0 | 53.1 | 63.8 | - | - | - | - |
| Vicuna-13B (Dir) | 50.3 | 28.0 | 39.6 | 52.8 | 64.3 | 53.6 | 67.1 | 63.8 |
| Vicuna-13B (Dir+Inta) | 55.5 | 36.0 | 43.0 | 56.8 | 63.8 | 49.0 | 62.0 | 63.6 |
| Vicuna-13B (TS) | 57.1 | 44.6 | 57.3 | 56.5 | 68.3 | 55.6 | 69.3 | 66.5 |
| Vicuna-13B (TS+Inta) | 55.1 | 35.6 | 49.3 | 52.1 | 66.6 | 56.3 | 66.3 | 64.0 |
| Vicuna-13B (ST) | 54.6 | 33.6 | 46.3 | 56.6 | 65.0 | 54.0 | 62.8 | 62.1 |
| Vicuna-13B (ST+Inta) | 53.6 | 37.0 | 44.3 | 55.1 | 69.5 | 57.3 | 67.6 | 66.3 |
| BLOOMZ-176B (Dir) | 53.5 | 38.3 | 44.3 | 54.3 | 63.1 | 51.0 | 61.0 | 63.8 |
| BLOOMZ-176B (TS) | 52.3 | 37.0 | 37.6 | 53.6 | 59.3 | 48.3 | 61.3 | 58.8 |
| BLOOMZ-176B (ST) | 54.3 | 37.3 | 44.3 | 55.5 | 59.5 | 48.0 | 60.3 | 60.3 |
| Davinci-003 (Dir) | 60.0 | 33.0 | 55.0 | 59.3 | 71.3 | 60.6 | 76.6 | 69.5 |
| Davinci-003 (TS) | 56.3 | 26.6 | 38.6 | 54.3 | 68.5 | 54.3 | 68.6 | 68.3 |
| Davinci-003 (ST) | 62.8 | 46.6 | 54.0 | 61.5 | 68.6 | 61.6 | 77.0 | 70.1 |
| ChatGPT (Dir) | 63.1 | 45.3 | **70.0** | 65.3 | 74.0 | 64.0 | 82.0 | 71.6 |
| ChatGPT (Dir+Inta) | 58.0 | 45.0 | 58.0 | 60.6 | 68.3 | 65.3 | 76.0 | 69.1 |
| ChatGPT (TS) | 63.0 | 49.6 | 59.0 | 62.8 | 71.5 | 62.0 | 77.6 | 71.8 |
| ChatGPT (TS+Inta) | 64.5 | 49.3 | 61.3 | 60.5 | 70.0 | 59.0 | 77.0 | 69.8 |
| ChatGPT (ST) | 64.3 | 51.6 | 64.0 | 62.3 | 72.3 | 63.6 | 77.0 | 74.0 |
| ChatGPT (ST+Inta) | 64.1 | 51.0 | 60.6 | 65.3 | 69.1 | 60.3 | 73.6 | 67.0 |
| GPT4 (Dir) | 64.0 | 48.6 | 67.6 | **67.0** | **75.3** | 68.0 | **83.6** | **74.5** |
| GPT4 (Dir+Inta) | 62.0 | 50.6 | 57.3 | 63.1 | 70.0 | **68.6** | 77.6 | 70.3 |
| GPT4 (TS) | **66.0** | **55.3** | 63.3 | 65.6 | 73.6 | **68.6** | 79.0 | 72.0 |
| GPT4 (TS+Inta) | 62.3 | 48.6 | 59.0 | 63.8 | 65.1 | 60.3 | 69.3 | 68.3 |
| GPT4 (ST) | 63.0 | 52.6 | 64.0 | 62.3 | 72.8 | 67.3 | 80.3 | 72.3 |
| GPT4 (ST+Inta) | 60.6 | 46.3 | 56.3 | 63.5 | 70.8 | 62.6 | 79.0 | 70.1 |

Table 4: Evaluation results (judged by ChatGPT) on XSAMSum. Pink denotes the fine-tuned baseline. Light blue and blue denote the zero-shot performance of open-source and non-open-source LLMs, respectively. "-" denotes the model cannot be evaluated in the corresponding dataset.

can easily realize which summaries are generated by LLMs or fine-tuned models. In this manner, the evaluators may have biases during scoring each summary. To ensure the fairness of human judgment, Stiennon et al. (2020) only retain the generated summaries whose length belongs to a certain range, and then collect human judgment on these summaries to minimize the potential evaluation bias caused by summary length. In our scene, the text styles of LLMs and fine-tuned models are quite different, which might also lead to bias. Thus, the human evaluation of comparing zero-shot LLMs and fine-tuned models on CLS needs more carefully designed.

As an alternative to human evaluation, recent studies (Liu et al., 2023; Kocmi and Federmann, 2023; Wang et al., 2023) show that the natural language generation (NLG) results evaluated by LLMs could achieve better correlations with humans. Following Wang et al. (2023), we utilize ChatGPT to score the generated summaries in a reference-free manner on four aspects, *i.e.*, coherence, relevance, consistency and fluency. An example prompt is shown in Figure 4, and please refer to Wang et al. (2023) for prompts of all aspects. The instruction of each aspect (marked in purple in Figure 4) is inspired by SummEval (a widely-used summarization meta-evaluation benchmark dataset) (Fabbri et al., 2021). We utilize the official

Score the following dialogue summarization given the corresponding dialogue with respect to fluency on a continuous scale from 0 to 100, where a score of zero means "disfluency" and score of one hundred means "perfect fluency". Note that fluency measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

Dialogue: [a given dialogue]
Summary: [one generated summary]
Scores:

Figure 4: An example prompt used to guide ChatGPT to score the summarization results in the aspect of fluency. Purple indicates the detailed instruction of the corresponding aspect.

APIs provided by OpenAI[12] to conduct the experiments with `gpt-3.5-turbo` model, and set the temperature to 0 to eliminate the randomness of evaluation results.

Table 4 shows the evaluation results on XSAMSum (En⇒Zh/De). As we can see, GPT-4 achieves the best performance in most aspects, showing its superiority. Besides, compared with the fine-tuned mBART-50 baseline, several zero-shot LLMs, including ChatGLM-6B, Vicuna-13B, Davinci-003, ChatGPT and GPT-4, achieve better results in all aspects, demonstrating the potentiality of performing zero-shot CLS via LLMs. For example, ChatGLM-6B (Dir+Inta) achieves 60.6, 35.3, 55.1 and 60.8 scores in aspects of coherence, relevance, consistency and fluency respectively on XSAMSum (En⇒Zh), while the counterparts of mBART-50 are 54.0, 32.3, 36.6 and 55.8, respectively. For GPT-4 (Dir), the corresponding scores even reach 64.0, 48.6, 67.6 and 67.0, significantly better than the fine-tuned mBART-50.

Moreover, while we show the interactive prompt can improve the performance of zero-shot LLM in terms of ROUGE scores and BERTScore, we do not find the same trend in the LLM-based evaluation results. In some cases, the interactive prompt even leads to worse LLM-based scores. We conjecture that the interactive prompt would force zero-shot LLMs to reduce the length of the generated summaries, and the models cannot make a good trade-off between conciseness and other aspects, which is also hard for humans.

[12] https://platform.openai.com/docs/guides/gpt/chat-completions-api

# 4 Related Work

## 4.1 Cross-Lingual Summarization

Given documents in one language, cross-lingual summarization (CLS) generates summaries in another language. Early work typically focuses on pipeline methods (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015), *i.e.*, translation and then summarization or summarization and then translation. Recently, with the availability of large-scale CLS datasets (Zhu et al., 2019; Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021; Wang et al., 2022a; Zheng et al., 2022), many researchers shift the research attention to end-to-end CLS models. According to a comprehensive CLS review (Wang et al., 2022b), the end-to-end models involve multi-task learning (Cao et al., 2020; Bai et al., 2021b; Liang et al., 2022b), knowledge distillation (Nguyen and Luu, 2022), resource-enhanced (Zhu et al., 2020; Jiang et al., 2022) and pre-training (Xu et al., 2020; Chi et al., 2021) frameworks. However, none of them explore LLMs performance on CLS. To our knowledge, we are the first to explore *can LLMs perform zero-shot CLS* and *how their results are*.

## 4.2 Large Language Models

Recently, there are growing interest in leveraging LLMs for various NLP tasks. Bang et al. (2023), Qin et al. (2023) and Zhong et al. (2023) conduct systematic investigations of ChatGPT's performance on various downstream tasks. Jiao et al. (2023) and Peng et al. (2023) evaluate ChatGPT on machine translation. Yong et al. (2023) show that ChatGPT could generate high-quality code-mixed text. Tan et al. (2023) explore the performance of ChatGPT on knowledge-based question answering (KBQA). Some works (Kocmi and Federmann, 2023; Wang et al., 2023; Liu et al., 2023; Ji et al., 2023) utilize ChatGPT or GPT-4 as an evaluation metric to assess task-specific model performance.

# 5 Conclusion and Future Work

In this technical report, we evaluate the zero-shot performance of mainstream bilingual and multilingual LLMs on cross-lingual summarization. We find that Davinci-003, ChatGPT and GPT-4 can combine the ability to summarize and translate to perform zero-shot CLS, and achieve competitive results with the fine-tuned baseline (*i.e.*, mBART-50). In addition, the current open-source LLMs (*i.e.*, BLOOMZ, ChatGLM-6B, Vicuna-13B and ChatYuan) generally show their limited ability to perform CLS in an end-to-end manner, showing the challenge of performing zero-shot CLS still exists.

In the future, we would like to unleash the potentiality of LLMs and leverage LLMs to perform CLS in few-shot learning manners.

## Limitations

While we evaluate the performance of LLMs on the cross-lingual summarization task, there are some limitations worth noting: (1) We only evaluate the lower threshold of these models' CLS performance. Prompts are important to guide LLMs to perform specific tasks, and future work could explore better prompts to obtain better results. (2) This report only uses two cross-lingual directions (En⇒Zh and En⇒De) in experiments, and all the languages are considered high-resource languages in the world. The performance of LLMs on low-resource languages still needs to be explored. According to Jiao et al. (2023), the machine translation ability of ChatGPT is limited on low-resource languages. We conjecture that the same situation might exist in CLS. (3) Though the general trend of the evaluation results should be correct, the comparisons between LLMs are not rigorous due to the decoding strategies of these models are not the same.[13] This is one of the major reasons leading to the limited soundness of this work. (4) In the future, we would like to conduct human evaluation to give more analyses.

## Acknowledgement

We thank anonymous reviewers for their constructive suggestions and comments.

## References

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Bai, Yang Gao, and Heyan Huang. 2021a. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924, Online. Association for Computational Linguistics.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. 2021b. Bridging the gap: Cross-lingual summarization with compression rate. *ArXiv preprint*, abs/2110.07936v1.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

[13]Currently, we cannot set the decoding strategy of GPT-4 when manually evaluating it on the ChatGPT platform. Besides, it is difficult to ensure the decoding strategies of LLMs are totally the same when using online platforms or demos.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *ArXiv preprint*, abs/2112.08804v1.

Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences. *arXiv preprint arXiv:2303.07610*.

Shuyu Jiang, Dengbiao Tu, Xingshu Chen, R. Tang, Wenxian Wang, and Haizhou Wang. 2022. Clue-GraphSum: Let key clues guide the cross-lingual abstractive summarization. *ArXiv preprint*, abs/2203.02797v2.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2:245–269.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022a. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022b. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2099, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022. Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6959–6969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11103–11111.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *Available at SSRN 4390455*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555, Portland, Oregon, USA. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference* *on Natural Language Processing*, pages 536–541, Suzhou, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Lisbon, Portugal. Association for Computational Linguistics.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, Yin Lin Tan, et al. 2023. Prompting large language models to generate code-mixed texts: The case of south east asian languages. *arXiv preprint arXiv:2303.13592*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Shaohui Zheng, Zhixu Li, Jiaan Wang, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022. Long-document cross-lingual summarization. *arXiv preprint arXiv:2212.00586*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.