Translating Movie Subtitles by Large Language Models using Movie-meta Information

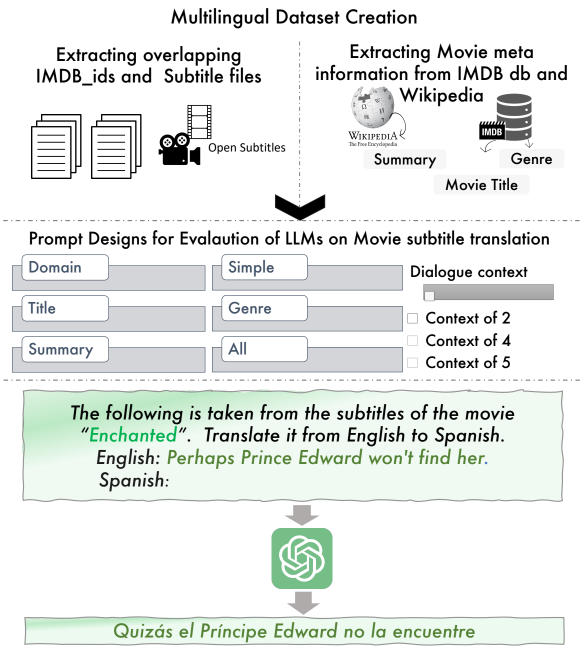
Ashmari Pramodya, Yusuke Sakai, Justin Vasselli,
Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST)
`{pussewala.ashmari.ow4, sakai.yusuke.sr9, vasselli.justin_ray.vk4,`
`kamigaito.h, taro}@is.naist.jp`

Abstract

Large language models (LLMs) have advanced natural language processing by understanding, generating, and manipulating texts. Although recent studies have shown that prompt engineering can reduce computational effort and potentially improve translation quality, prompt designs specific to different domains remain challenging. Besides, movie subtitle translation is particularly challenging and understudied, as it involves handling colloquial language, preserving cultural nuances, and requires contextual information such as the movie's theme and storyline to ensure accurate meaning. This study aims to fill this gap by focusing on the translation of movie subtitles through the use of prompting strategies that incorporate the movie's meta-information, e.g., movie title, summary, and genre. We build a multilingual dataset which aligns the OpenSubtitles dataset with their corresponding Wikipedia articles and investigate different prompts and their effect on translation performance. Our experiments with GPT-3.5, GPT-4o, and LLaMA-3 models have shown that the presence of meta-information improves translation accuracy. These findings further emphasize the importance of designing appropriate prompts and highlight the potential of LLMs to enhance subtitle translation quality.

1 Introduction

Large language models (LLMs) trained on large unlabeled corpora have emerged as powerful tools in the field of natural language processing (NLP) (Zhao et al., 2025) under model scaling, which allows prompting for downstream applications (Chowdhery et al., 2023; Brown et al., 2020; Laskar et al., 2023). As a result, a new paradigm of pretrain, prompt, and predict has emerged (Liu et al., 2023), enabling LLMs to perform very high-quality machine translation (MT), even though they were not explicitly trained for this task (Brown et al., 2020). While studies on prompting for MT exist (Zhang et al., 2023; Puduppully et al., 2023), the application of LLMs across different MT domains (Eschbach-Dymanus et al., 2024) still presents opportunities for further exploration.


Multilingual Dataset Creation
Extracting overlapping IMDB_ids and Subtitle files
Open Subtitles
Extracting Movie meta information from IMDB db and Wikipedia
Wikipedia
IMDB
Summary
Genre
Movie Title
Prompt Designs for Evalaution of LLMs on Movie sutbtitle translation
Domain
Simple
Title
Genre
Summary
All
Dialogue context
Context of 2
Context of 4
Context of 5

The following is taken from the subtitles of the movie "Enchanted". Translate it from English to Spanish.
English: Perhaps Prince Edward won't find her.
Spanish:

Quizás el Príncipe Edward no la encuentre

Figure 1: Quick overview of the Multilingual dataset creation process and the Prompt design for evaluating LLMs.

In this paper, we focus on prompting LLMs for MT, specifically targeting the translation of movie subtitles. In machine translation, translating subtitles poses particular challenges due to accuracy and context sensitivity (Karakanta et al., 2022). Movie subtitle translation requires the disambiguation of polysemous terms, e.g., "chamber", based on the context provided by the story and scenes and also the handling of colloquial phrases and slang (Gupta et al., 2019). This study aims to address these challenges by integrating the movie's meta-information, such as the title, genre, summary, and categories, into the translation prompt and evaluating how the

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 315–330
July 28-29, 2025 ©2025 Association for Computational Linguistics

performance of LLMs varies.

We create a multilingual, context-enriched dataset by mapping subtitles to corresponding movie meta-information, where the title and genre are sourced from IMDb, and the summary is obtained from Wikipedia. This dataset[1] focuses on translations from English into four languages: German, Spanish, French, and Finnish. We evaluate various prompting strategies for LLMs that integrate this meta-information to improve subtitle translation, using GPT-3.5, GPT-4o, and LLaMA-3 as testbeds. We aim to compare the effects of different types of movie meta-information, e.g., title, summary, and genre, on translation accuracy to understand how these elements influence the quality of translations, as shown in Figure 1.

Our findings shows that 1) while meta-information does not drastically change translation quality, including the *movie title* consistently improves performance, with GPT-4o seeing the greatest improvement. 2) LLaMA-3 struggles with complex contextual information, such as summaries. 3) Including previous dialogue lines improves the scores compared to simple prompts. 4) Combining meta-information with dialogue context yields strong results, especially for LLaMA-3, although the overall improvements remain modest. 5) Spanish (En-Es) benefited most from the additional information. These findings highlight the importance of prompt design in improving subtitle translation quality.

2 Background and Related Work

2.1 Prompt Engineering

Prompt engineering is the process of creating a suitable prompt that gets the best performance on the downstream task (Patel et al., 2023). In general, there are four major factors that guide the LLMs in performing tasks effectively: the task description, input data, contextual details, and prompt style (Zhao et al., 2025). Therefore effectiveness of prompting is highly influenced by how the prompt is presented, with even minor changes potentially leading to differences in performance. This has motivated researchers to create more advanced prompting techniques to maximize the potential of LLMs. Previous studies have found that LLMs can perform machine translation without being specifically fine-tuned (Radford et al., 2019).

2.2 Translations by LLMs

Finding the right prompt recipe to enhance MT accuracy with LLMs has become a topic of research (Zhang et al., 2023). Most research has focused on using simple prompts like `{Source text} = {Target text}` or `Translate to {language_name} :{text}` (Brown et al., 2020; Zhang et al., 2023). Moslem et al. (2023) examined GPT-3 and GPT-3.5 for MT, focusing on domain-specific adaptation, while Bawden and Yvon (2023) found they often fall short of SOTA MT systems and commercial translators.

Briakou et al. (2023) studied the impact of LLM data on MT. Recently, Vilar et al. (2023), investigated the use of prompting with PaLM (Chowdhery et al., 2023) for translation and found that even randomly selected high-quality examples can perform as well as or better than those chosen based on input relevance. Agrawal et al. (2023) explored input-specific examples and found that n-gram overlap enhances prompt effectiveness.

A comprehensive study of how different prompting strategies influence performance was lacking. So, a case study was done by Zhang et al. (2023) focusing on GLM-130B (Zeng et al., 2023) and found that prompting performance varies widely across different templates, with simple English templates generally working best for machine translation, and language-specific templates are effective when translating into languages the LLMs were pre-trained on. Inspired by the human translation process, He et al. (2024) proposed MAPS, which involves three steps: knowledge mining, knowledge integration, and knowledge selection. Evaluation on the WMT22 test set shows that MAPS improves the performance of models like text-davinci-003 and Alpaca.

Despite these advancements, Zhang et al. (2023) point out that prompting for machine translation still faces challenges such as copying errors, mistranslation of entities, hallucinations, poor direct translation between non-English languages, and the "prompt trap," where translating the prompt itself becomes complex and problematic.

2.3 Subtitle Translation

Recent research shows that Neural Machine Translation (NMT) can be highly effective for movie subtitle translation, especially with post-editing to reduce effort (Huang and Wang, 2023). However, challenges including subtitle block limitations, lex-

---

[1] https://huggingface.co/datasets/Ash96/SubtitleMetaData

ical consistency, lexical errors such as the translation of idioms and figurative language, and context-related errors persist (Karakanta et al., 2022).

3 Prompting for MT with Meta-information

3.1 Dataset Creation

For the multilingual dataset, we selected the language pairs from English to French, German, Spanish, and Finnish in OpenSubtitles 2018 (Lison et al., 2018). The OpenSubtitles dataset is a large collection of parallel corpora containing multilingual subtitles from movies and TV shows. It is freely available to the research community on the OPUS[2]. These particular language pairs were selected because they are well supported by LLMs and also share the same Latin script. We included Finnish because it is both a gender-neutral and agglutinative language, whereas Spanish, German, and French are gendered and fusional languages.

To create the dataset, we first downloaded XML files from the OPUS website. Each file contains subtitles for a specific language pair and includes meta-data about the subtitle and its associated movie or TV episode, such as the title, release year, and IMDb identifier in numerical format. Here, IMDb (Internet Movie Database) [3] is an online platform that provides detailed information about movies, TV shows, actors, and production details.

These files encoded information using the format `lang/year/imdb_id/opensubtitles_id.xml.gz`, where sentence IDs align across languages. Following discussions on the Hugging Face GitHub[4] and using the script from HuggingFace[5], we combined the data into a JSON format. This included meta-data like IMDb ID, subtitle ID, sentence ID, and translations across parallel files for each language pair.

Next, we extracted overlapping IMDb IDs to obtain subtitle files for the same movie across languages, followed by aligning the overlapping subtitle IDs with English sentence IDs to ensure consistency across languages. Meta-data such as movie titles and genres were sourced from the IMDb database, and movie summaries were retrieved from Wikipedia articles in all language pairs. The dataset consist of 10,777 and 21,575 parallel sentences for testing and training, respectively, across the four languages. The statistics of the datasets are provided in the Table 9 and 10 in Appendix A.

3.2 Prompting Strategy for MT

We designed the zero-shot prompts, which were structured mainly around two components: meta-information integration and contextual integration.

Meta-Information Integration We designed a total of six prompt templates as shown in Table 1. The first prompt *simple* is a simple template from Zhang et al. (2023), and the second prompt *movie domain* serves as the base template for our study which includes the domain information of movie subtitles. The following prompts were derived from it to include specific meta-information: *title* (movie title), *summary* (movie summary), *genre* (movie genre), and *all* which incorporates all three.

Contextual Integration As shown in Table 2, we designed the prompts to include the previous N lines (N=2 to N=5) as dialogue context to measure the impact of prior contexts without meta-information (Rikters et al., 2021).

Combining Meta-Information and Contextual Integration Our preliminary studies show that the best-performing meta-information prompt was "*title*" and the optimal context length without meta-information was 4. Therefore, we selected N=4 from the previous dialogue line evaluations, combined with the title, to further enhance translation quality. This approach aimed to leverage both the focused context provided by the movie title and the conversational flow from preceding lines, assessing whether this combination produced better results than using either method alone.

4 Experimental Setup

In this section, we outline the experimental setup used to evaluate the impact of different prompting strategies on subtitle translation quality. We compare various levels of meta-information, including movie titles, summaries, and genres, using the OpenSubtitles dataset across multiple language pairs using LLaMA-3 GPT-3.5 and GPT-4o. We also examine the effect of incorporating previous dialogue context to enhance translation accuracy. In addition, we compared our method against the MAPS framework (He et al., 2024).

---

[2] https://opus.nlpl.eu/

[3] https://www.imdb.com/

[4] https://github.com/huggingface/datasets/issues/1844

[5] https://huggingface.co/datasets/Helsinki-NLP/open_subtitles/blob/main/open_subtitles.py

ID	Template (in English)

simple	`English: {en_sentence}`
	`[tgt] :`

movie domain	`The following is taken from the subtitles of a movie. Translate it from English to [tgt]`
	`English: {en_sentence}`
	`[tgt]:`

title	`The following is taken from the subtitles of the movie {title}. Translate it from English to [tgt]`
	`English: {en_sentence}`
	`[tgt]:`

summary	`Here is a summary of a movie: {summary}`
	`The following is taken from the subtitles of that movie. Translate it from English to [tgt]`
	`English: {en_sentence}`
	`[tgt]`

genre	`The following is taken from the subtitles of the {genre} movie. Translate it from English to [tgt]`
	`English: {en_sentence}`
	`[tgt]:`

all	`Here is a summary of the {genre} movie {title}: {summary}`
	`The following is taken from the subtitles of that movie. Translate it from English to [tgt]`
	`English: {en_sentence}`
	`[tgt]`

Table 1: Templates for translation prompts incorporating meta-information. The target language name is represented by the `tgt` while `en_sentence` represents the source text, which is a subtitle.

4.1 LLM Models

For evaluation, we used leading LLMs alongside traditional NMT systems. We examined Meta's LLaMA-3-70B-Instruct (Grattafiori et al., 2024), GPT-3.5-turbo-0125 (Brown et al., 2020), GPT-4o-2024-05-13 (OpenAI et al., 2024), and the multilingual NMT model M2M100 (Fan et al., 2021).

4.2 Evaluation Metrics

Automatic Evaluation We adopted the widely used COMET score (Rei et al., 2020) as our primary evaluation metrics. Additionally, BLEU score (Papineni et al., 2002) and chrF++ (Popović, 2017) were used. BLEU and chrF++ focus on surface-level features by comparing the n-grams, while COMET is a neural network-based metric that captures semantic meaning more effectively. Furthermore, statistical significance testing (Koehn, 2004) was performed using SacreBLEU (Post, 2018) with the default parameters for significance testing with paired bootstrap resampling, where $p < 0.05$ means the difference is significant.

Human Evaluation In addition to automatic evaluations, we conducted a human evaluation to better understand the impact of incorporating meta-information. This is based on relative ranking (Callison-Burch et al., 2008), a method commonly used in WMT tasks, where translations are ranked relative to each other. Native speakers were used as annotators for each language, with two annotators assigned to each language except Finnish, where no annotators were available. Each annotator was given all the sentences to rank from best to worst. For this task, we selected 40 entries which had six distinct translation outputs from the GPT-4o model.

```
Here is a dialogue taken from a movie, translate the
last line from English to [tgt].

Line 1
Line 2
⋮
Line N

English: {en_sentence}
[tgt]
```

Table 2: Translation prompts using previous context. The target language name is represented by the `tgt` while `en_sentence` represents the source text, which is a subtitle. In this setup, we consider the number of previous sentences, ranging from N=2 to N=5.

5 Results

Table 3 summarizes the performance across different language pairs based on the prompting strategies detailed in Section 3.2.

Domain Knowledge Although prior studies (Zhang et al., 2023) show that *simple* prompts obtain good results in general, it is simply outperformed by "*movie domain*" which explicitly includes the domain knowledge of movies. This small amount of additional domain information generally leads to slight improvements in translation quality over *simple*, resulting in modest increases in BLEU and COMET scores across most language pairs. For example, En→Es direction achieves gains of 1.07 BLEU points with GPT-4o. However, performance drops were observed in the En→Fi direction with GPT-3.5 with 0.78 BLEU points. Although still relatively simple, this prompt helps the model recognize that the task involves translating movie subtitles, which can aid in understanding colloquial language, idiomatic expressions, and cultural references typical of film scripts. By explicitly indicating that the input is a movie subtitle, the models are better equipped to make informed translation choices with a significant difference.

Models	Template ID	En→Es		En→De		En→Fr		En→Fi	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
M2M100	–	0.7902	21.7	0.7502	18.0	0.7906	17.4	0.7906	11.8
MAPS$_{\text{LLaMA-3 COMET}}$	He et al. (2024)	0.8230	24.97	0.8060	20.57	0.7830	19.96	0.8260	13.69
GPT-4o	simple	0.8484	32.62	0.8231	26.43	0.7638	26.67	0.8685	19.91
	movie domain	0.8523	33.69	0.8253	26.83	0.8072	26.50	0.8712	20.59
	+ N = 2	0.8518	33.91	0.8265	27.20	0.8057	26.67	0.8717	20.64
	+ N = 3	0.8521	33.92	0.8268	27.17	0.8064	26.79	0.8716	20.67
	+ N = 4	0.8522	34.03†	0.8272	27.12	0.8065	26.74	0.8718	20.83
	+ N = 5	0.8510	33.97	0.8262	27.25	0.8065	26.85	0.8267	20.74
	+ title	**0.8540**	34.01	**0.8280**	**27.33**†	**0.8079**	26.23	**0.8724**	20.81
	+ summary	0.8522	33.96	0.8252	27.30	0.8074	**26.96**†	0.8723	**20.92**†
	+ genre	0.8521	33.96	0.8269	27.08	0.8074	26.62	0.8719	20.62
	all	0.8527	**34.26**†	0.8259	27.29	0.8072	26.88	0.8721	20.86
	title + N = 4	0.8543	34.06	0.8278	27.34†	0.8082	26.93	0.8727	20.93†
GPT-3.5	simple	0.8472	33.01	0.8206	26.01	0.8012	26.01	0.8607	**20.04**
	movie domain	0.8493	33.02	0.8224	26.07	0.8023	25.96	0.8626	19.29
	+ N = 2	0.8474	32.93	0.8186	25.96	0.8007	26.14	0.8582	19.20
	+ N = 3	0.8493	33.14	0.8216	26.13	0.8020	26.21	0.8603	19.27
	+ N = 4	0.8494	33.15	0.8219	26.17	0.8027	26.31	0.8623	19.41
	+ N = 5	0.8328	33.15	0.8214	26.14	0.8028	26.29	0.8618	19.32
	+ title	**0.8500**	33.19	**0.8233**	**26.28**	**0.8036**	**26.23**	**0.9763**	19.29
	+ summary	0.8099	**34.25**†	0.8232	25.92	0.8019	26.16	0.8609	19.28
	+ genre	0.8491	33.01	0.8229	26.16	0.8022	26.03	0.8618	19.27
	all	0.8328	29.40	0.8230	25.89	0.8019	26.05	0.8613	19.17
	title + N = 4	0.8495	33.29	0.8227	26.24	0.8034	26.34†	0.8626	19.45
LLaMA-3	simple	0.8202	29.57	0.8077	24.22	0.7850	23.14	0.8232	14.65
	movie domain	0.8354	29.67	0.8119	24.05	0.7876	23.07	0.8349	15.60
	+ N = 2	0.8367	29.94	0.8109	24.13	0.7896	23.88	0.8307	15.15
	+ N = 3	0.8368	29.98	0.8113	24.23	0.7893	23.76	0.8307	15.19
	+ N = 4	0.8369	29.99†	0.8113	24.33	0.7894	23.79	0.8308	15.27
	+ N = 5	0.8365	29.93	0.8111	24.24	0.7892	23.87 †	0.8300	15.16
	+ title	**0.8360**	**29.72**	**0.8137**	**24.39**	**0.7897**	**23.21**	**0.8351**	**15.66**
	+ summary	0.8291	29.64	0.8077	24.13	0.7591	23.20	0.8042	15.64
	+ genre	0.8354	29.59	0.8109	24.05	0.7889	23.01	0.8335	15.61
	all	0.8310	29.64	0.8093	24.23	0.7572	22.97	0.8293	15.54
	title + N = 4	0.8377	30.09†	0.8121	24.45	0.7902	23.88†	0.8309	15.24

Table 3: COMET and BLEU scores for zero-shot prompts including meta-information and previous context for GPT-3.5, GPT-4o, and LLaMA-3 models. The rows labeled N=2 to N=5 show the results of using previous context lines in the prompt. The highest scores for meta-information are in bold, while the highest scores for context are underlined. Cells highlighted in red indicate the overall highest scores across both meta-information and context. Moreover, the decoration of † on the best scores for each section means it is significantly different than baselines according to the significance test with $p < 0.05$.

Contextual Integration Including previous lines as context generally improves translation quality across all language pairs over the *simple* and *movie domain* prompts. For instance, En→Es using GPT-4o sees a slight increase in BLEU from 32.62 to 34.03 and a considerable gain in COMET from 0.8253 to 0.8522 when 4 lines of previous context are added. For most language pairs, N=4 appears to be the optimal number of previous context lines, providing the best balance between translation accuracy and context usage.

Meta Information Incorporating meta-information such as *title*, *summary*, and *genre* into the prompts enhances the quality of translation in all metrics over the baselines. The use of *title* consistently improves translation performance with modest gains in COMET scores in all language pairs. This trend is noticeable in all models, but especially in GPT-3.5 and LLaMA-3, where the inclusion of movie *title* improves BLEU scores in En→De, En→Es and En→Fr language pairs. Compared with GPT-4o's BLEU results, the improvements are especially clear in En→De direction with a gain of 0.9 BLEU points.

summary yields mixed results, with slight BLEU gains for En→Fr and En→Fi using GPT-4o but lower COMET scores than *title* (gaining 0.73 and 0.11 BLEU points, respectively, over the *title*); however, the difference is not significant. In fact, with LLaMA-3, the performance of the *summary* is lower than the *title* for all language pairs. The decrease in performance observed when using summaries as context can be attributed to the increased cognitive load associated with processing longer prompts. On average, summaries contain approximately 980 tokens, compared to the significantly shorter length of titles, which average around 60 tokens. This disparity in input length likely overwhelms the model, diverting its focus from the essential information needed for accurate translation. These findings align with prior research by Levy et al. (2024), which showcases how longer input sequences can impact the reasoning performance of LLMs.

The *genre* prompt produces variable results and is often less effective than the *title* prompts. This may be because genre does not provide as direct a context as the title, resulting in less improvements. The *all* prompt shows moderate improvement in both BLEU and COMET, though it does not exceed the performance of the *title* prompt. However, for the En→Es language pair it performs better than *title* especially with GPT-4o, where it ranks the highest among all prompts. In contrast, GPT-3.5 shows a significant drop, with a decrease of 3.91 BLEU points, which is much lower than the other prompts. This may be due to the limited capacity to handle multiple pieces of information effectively in GPT-3.5. For other language pairs, the *all* prompt does not perform well. While it gives detailed context, using too many meta-information elements can make things too complicated, leading to a drop in translation quality. However, in LLaMA-3, adding meta-information does not perform better than using just the previous context. Overall, GPT-4o performed best among all models. The En→Es direction achieved the highest BLEU score, while En→Fi had lower BLEU but higher COMET scores due to Finnish's agglutinative nature, making word-for-word matches challenging.

prompt type	En→Es	En→De	En→Fr
simple	0.510	0.553	0.455
movie domain	0.577	0.553	**0.615**
+ title	**0.593**	**0.600**	0.565
+ summary	0.493	0.340	0.525
+ genre	0.397	0.500	0.400
all	0.430	0.453	0.440

Table 4: Expected wins for different prompt types across language pairs in human evaluation task

Combining Meta-Information and Contextual Integration This shows greater gains, particularly in GPT-4o and LLaMA-3. For example, in the En→Fi direction with GPT-4o, the BLEU score improves by 1.02 over the simple prompt, and the COMET score increases from 0.8675 to 0.8727. GPT-3.5 sees moderate improvement, but performs better with just meta-information. LLaMA-3 benefits the most, especially in the En→Es direction, where the BLEU score increases from 29.57 to 30.09, with a statistically significant difference, and the COMET score increases from 0.8202 to 0.8377.

We also evaluated the MAPS framework (He et al., 2024) using LLaMA-3 model on our test dataset and observed that our method achieves higher scores in subtitle translation. Although MAPS effectively integrates external knowledge for context-rich tasks, it is less effective for subtitles, which are fragmented and lack sufficient context, limiting the usefulness of the mined knowledge. In contrast, our approach leverages the unique characteristics of subtitles, such as their brevity and conversational tone, to deliver more accurate and contextually appropriate translations.

Human Evaluation Table 4 shows the summary of Expected Wins, which computes the probability that the system's translation is ranked higher compared to a randomly chosen opposing system, evaluated on a randomly selected sentence by a randomly picked judge (Bojar et al., 2014). A higher score indicates a better performance in human evaluation. For En→Es and En→De, the probability

Metric	Shot	En→Es	En→De	En→Fr	En→Fi
BLEU	0-shot	29.72	24.39	23.21	15.66
	3-shot	30.72	24.86	24.54	15.95
	5-shot	31.19	25.09	24.72	16.10
COMET	0-shot	0.8360	0.8137	0.7897	0.8357
	3-shot	0.8395	0.8133	0.7915	0.8380
	5-shot	0.8413	0.8149	0.7921	0.8395
chrF++	0-shot	56.72	51.63	50.24	46.86
	3-shot	56.84	51.52	50.56	47.21
	5-shot	57.22	51.73	50.56	47.27

Table 5: Few-shot learning results on LLaMA-3

of a sentence being translated accurately is higher with *title*, making it the most effective for these language pairs. In contrast, for En→Fr, the *movie domain* yields the best performance. These results suggest that adding meta-information, such as a summary, does not necessarily improve translation accuracy. The consistency in scores between *simple* and *movie domain* for En→De indicates that both prompts are equally effective for this language pair, with a higher likelihood of accurate translation without the need for complex meta-information.

Few-shot Learning We evaluate the few-shot learning performance of LLMs. Few-shot learning is also denoted as K-shot, with K representing the number of examples provided before the query, where in our case, examples are randomly sampled from the training set. For this we used the prompt title detailed in Appendix section B.3, as our earlier results showed that movie titles provide a strong signal for subtitle translation, while summaries or genres may introduce noise due to varying levels of detail. The experiment results are presented in Table 5. When K >= 3, the model consistently outperforms the 0-shot scenarios. This indicates that few-shot prompting clearly improves translation quality by leveraging the provided examples.

6 Analysis

The experiment was initially designed based on the hypothesis that summaries would enhance subtitle translation quality more than titles due to their more detailed nature. However, the results revealed that prompts that included titles performed slightly better than those that included summaries. Although we expected a performance improvement with summaries, the difference in performance between the use of titles and summaries, measured by COMET and BLEU scores, was minimal. This suggests that while summaries provide more information, titles offer more focused and relevant context for subtitle translation.

	BERTScore	
Movie name	GPT-4o	LLaMA-3
The Chronicles of Narnia: Prince Caspian	0.8435	0.8281
Enchanted	0.8213	0.8319
The Duchess	0.8275	0.8090
Frozen Fever	0.8259	0.8352
Dreamgirls	0.8274	0.8063
The Life Before Her Eyes	0.8261	0.8264
High School Musical 2	0.8324	0.8309
Star Trek	0.8068	0.8012
Spider-Man 3	0.8229	0.8048
The Princess and the Frog	0.8319	0.8453
Thor	0.8335	0.8299
Dear John	0.8327	0.8399
Letters to Juliet	0.8506	0.8330
Gridiron Gang	0.8318	0.8184

Table 6: BERT Scores against the LLM generated summary to the Wikipedia summary.

Evaluating LLMs' Knowledge of Movie Plot Summaries: To investigate why including the *title* in the prompt performs better than including summaries, we conducted an experiment to check whether the content of a movie might already be familiar to LLMs when only the title is provided. This approach tested the hypothesis that LLMs, pre-trained on massive datasets, are able to retrieve accurate movie knowledge based on titles alone and leading to more effective subtitle translations.

To achieve this, we queried the LLM to generate plot summaries for each movie listed in Table 9 using the prompt *"What is the summary of the plot of this "title" movie?"*. Then, the responses generated by the models were compared to Wikipedia summaries to evaluate how accurately the LLMs could retrieve relevant pre-learned knowledge based solely on the movie titles.

Based on the results in Table 6, the high BERTScores (Zhang et al., 2020) show that the generated plot summaries are contextually similar to those found in Wikipedia. This suggests that the movie title alone provides sufficient information about the movie, likely because the model has been pre-trained on extensive sources, including Wikipedia. Using the title simplifies the prompt, allowing the model to leverage its pre-existing knowledge efficiently. These findings show that titles serve as short cues, allowing LLMs to retrieve more focused and relevant context for subtitle translation.

Instruction: `You know the following movie from your training data. What is the name that fills in the [MASK] token? The name is exactly one word long, and is not a pronoun or any other word. You must make a guess even if you are uncertain.`
Example:
`Input: The door opened, and [MASK], dressed and hatted, entered with a cup of tea.`
`Output: Gerty`

Input: `These are not the issues that burden the Duke, Lady [MASK].`

Table 7: Example of the name-cloze task in subtitles, where the model predicts a masked character name based on subtitle context alone.

Assessing LLMs' Subtitle Knowledge In addition to querying plot summaries, we evaluated whether the LLMs had prior exposure to subtitle data by asking them to predict the next sentence in a subtitle sequence using the prompt, "*Here is a subtitle from the movie title. Please provide the next sentence.*" We aimed for evidence of the ability of the models to memorize specific details of subtitles. Instead, they often produced generic responses indicating their inability to provide the specific line.

Therefore, we used the name-cloze method described by Chang et al. (2023) instead of predicting the next subtitle. Their method involves giving a passage from a book with a masked character name to the model and asking it to predict the masked word. This method helps evaluate the model's ability to recall and predict specific entities from the text. We applied this to 100 subtitles, each with a single proper entity, masking the name without providing the movie title as shown in Table 7. The model's name-cloze accuracy was only 3%, indicating that the context alone provided little information to infer the correct character name. However, when the title was included, the accuracy increased to 26%. This indicates that the title alone contains embedded information about the movie, providing enough context for the model to more accurately identify character names when the title is provided. This suggests that while LLMs may have broad movie knowledge from sources like Wikipedia, specific subtitle data is less accessible, and titles play a more significant role in aiding subtitle translation tasks.

Qualitative Analysis We used the PIE corpus (Adewumi et al., 2022) to evaluate idiomatic translation quality, extracting 20 idioms from the dataset. Spanish translations, generated using the title prompt, were assessed on a 1–3 scale (Li et al., 2024), with GPT-4o scoring 2.5 and LLaMA-3 scoring 2.4. Both models captured figurative meanings but often relied on literal or descriptive translations, indicating room for improvement in cultural nuance. In the Table 8 the title prompt ("À plus, mon frère") effectively captures both the idiomatic farewell ("Catch you on the fly") and the slang term ("homey") by using "À plus" (see you later) and "mon frère" (bro). In contrast, the simple prompt ("Attrape toi en vol, mon pote") translates the idiom too literally, while M2M ("Tu t'as pris dans le vol, Homéy.") is incorrect and misinterprets both the idiom and slang. Further analysis of the idiomatic and colloquialisms is provided in Appendix C.

English: French:	Catch you on the fly, homey. À plus, mon pote
M2M	Tu t'as pris dans le vol, Homéy.
simple	Attrape toi en vol, mon pote.
title	À plus, mon frère

Table 8: Example of a translation from English to French, including an idiomatic expression, generated by LLaMA-3.

7 Conclusion

In this work, we compare the performance of GPT-4o, GPT-3.5, and LLaMA-3 in translating movie subtitles, with a focus on how different types of meta-information, such as movie titles, summaries, and genres, impacted translation quality. Our results show that GPT-4o always outperformed the others for multiple language pairs, especially when movie titles were given in the prompt. Spanish translations (En→Es) benefited the most from additional context, while Finnish translations (En→Fi) posed challenges, with minimal gains from meta-information. Simpler prompts often led to more stable results, with basic prompts ranking higher in human evaluations.

Overall, this research shows the importance of prompt design in subtitle translation by LLMs, while meta-information can be useful in particular contexts, careful selection is essential in order not to fall into diminishing returns. Future work could explore testing the model's ability with low-resource languages to assess its performance in more challenging translation scenarios.

Limitations

This study has several limitations:

Human Evaluation First, due to resource constraints, human evaluation was not conducted for the English-Finnish (En→Fi) language pair, restricting a comprehensive assessment of this model's performance in that language.

Linguistic Analysis of Polysemy Another limitation in the study is that we did not conduct a linguistic analysis to evaluate how the subtitle translations handled polysemous words. Instead, we relied on BLEU and COMET scores and focused heavily on the impact of meta-information, such as movie titles, summaries, and genres, on translation quality.

Language Selection The study is limited by the selection of languages, and a broader evaluation across more diverse language pairs is necessary to better understand the model's capabilities across different linguistic contexts.

Evaluation Another limitation is our evaluation does not account for discourse-level effects of meta-information. Future work should explore discourse-aware metrics like APT (Miculicich Werlen and Popescu-Belis, 2017), and BlonDe (Jiang et al., 2022) to better capture phenomena such as pronoun translation and lexical consistency.

Knowledge Cutoff A further limitation is that the models lack awareness of movies released after their knowledge cutoff dates December 2023 for LLaMA-3[6], October 2023 for GPT-4o[7], and September 2021 for GPT-3.5. To address this, the method can be adapted for newly released movies by fine-tuning the model with additional training data collected. This approach would enable the model to incorporate updated domain knowledge and effectively handle subtitle translation for newly released movies. However, this approach faces challenges such as knowledge editing, which involves modifying specific information without extensive retraining, and continual learning, which ensures new information is integrated without causing catastrophic forgetting of previously learned knowledge (Ghosh et al., 2024).

Ethics Statement

In conducting this research, we adhered to ethical guidelines throughout the study. All data used, including subtitle translations and meta-information, was sourced from publicly available datasets (e.g., OpenSubtitles, IMDb, and Wikipedia). No personal or sensitive data was involved in the research process, ensuring privacy and data protection standards were met. Moreover, there is no harmful content included in the examples used in the paper. Additionally, human evaluations were conducted with full consent of the annotators. All recruited annotators were paid above the minimum wage.

References

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696, Marseille, France. European Language Resources Association.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

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

---

[6] `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`

[7] `https://platform.openai.com/docs/models/gpt-4o#gpt-4o`

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of LLM domain adaptation for business IT machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622, Sheffield, UK. European Association for Machine Translation (EAMT).

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. 2024. A closer look at the limitations of instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15559–15589. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Prabhakar Gupta, Mayank Sharma, Kartik Pitale, and Keshav Kumar. 2019. Problems with automating translation of movie/tv show subtitles. *Preprint*, arXiv:1909.05362.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Jie Huang and Jianhua Wang. 2023. Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators' effort. *Perspectives*, 31(4):620–640.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models.

Proceedings of the AAAI Conference on Artificial Intelligence, 38(17):18554–18563.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2023. Bidirectional language models are also few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy Chen. 2023. DecoMT: Decomposed prompting for machine translation between related languages using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4602, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2021. Japanese–english conversation parallel corpus for promoting context-aware machine translation research. *Journal of Natural Language Processing*, 28(2):380–403.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Dataset Statistics

The test dataset comprises 14 movie files, containing a total of 10,777 parallel sentences across the four languages, as shown in Table 9. In contrast, the training dataset consists of 20 subtitle files, 21,575 parallel sentences, as detailed in Table 10. Figure 2 presents a part of the collected data with meta information and translations.

```
[
  {
    "meta": {
        "imdb_id": "499448",
        "title": "The Chronicles of Narnia:
        Prince Caspian",
        "year": 2008,
        "genres": [
            "Action",
            "Adventure",
            "Family",
            "Fantasy"
        ],
        "summary": {
               "en":EN_SUMMARY
               "es":ES_SUMMARY
               "de":DE_SUMMARY
               "fr":FR_SUMMARY
               "fi":FI_SUMMARY
        }
     },
    "translations": [
        {

          "en_sentence_id": "3",
          "en": "You have a son.",
          "es": "Tenéis un hijo.",
          "fr": "Vous avez un fils.",
          "de": "Ihr habt einen Sohn.",
          "fi": "Teillä on poika."
        },
        {
          "en_sentence_id": "4",
          "en": "The heavens have blessed us.",
          "es": "Los cielos nos han bendecido.",
          "fr": "Les dieux nous ont bénis.",
          "de": "Der Himmel hat uns gesegnet.",
          "fi": "Taivas on siunannut meitä."
        },
        {  .
           .
         },
      ]
    },
    {


    }
]
```

Figure 2: Sample of the collected data: JSON structure containing movie meta-information and translations

B Experiments

B.1 Details of experiment settings

For the experiments, we used GPT-3.5-turbo-0125 (Brown et al., 2020) and GPT-4o-2024-05-13 (OpenAI et al., 2024), with the top_p set to 0 and temperature set to 0.5 for both models We also used Meta's LLaMA 3 (Grattafiori et al., 2024) for the experiments, conducted on a single NVIDIA RTX 6000 Ada GPU, with 4-bit quantization utilized for model generation.

Table 11 present the chrf++ scores across different language pairs based on the prompting strategies detailed in Section 3.2.

B.2 Additional Experiments

To verify the observed tendency, we collected 20 additional film files and tested them using the LLaMA-3 model with our methodology. The results in Table 12 indicate that the tendency remains consistent for the title.

B.3 Few shot Learning

The prompt template used is detailed in Table 13. Few-shot learning is also denoted as K-shot, with K representing the number of examples provided before the query, where in our case, examples are randomly sampled from the training set.

C Qualitative Analysis of Results

Colloquialisms and idioms are language constructs that depend upon culturally learned and contextually learned meanings to carry meaning beyond their literal expressions. In subtitle translation, these elements are particularly challenging as they must be concise while maintaining the original intent, tone, and cultural relevance.

Idioms To assess the translation quality of idiomatic expressions, we used the PIE corpus (Adewumi et al., 2022), which contains 1,197 idioms and 5,170 related sentences. From this dataset, we extracted English idioms that overlapped with those present in our movie subtitle dataset, resulting in a subset of 20 sentences. These idioms fell under the categories of personification, metaphor, and simile as classified in the PIE dataset, and were evaluated based on their Spanish translations generated using the title prompt. To evaluate idiom quality, we used the scoring method of Li et al. (2024), which assigns points from 1 to 3, where 1 indicates a completely inaccurate meaning, 2 suggests the meaning requires minor refinements, and 3 reflects a perfect capture of nuanced cultural meanings. We applied this method to the translations produced by LLaMA-3 and GPT-4o, with GPT-4o achieving an average score of 2.5, while LLaMA-3 scored 2.4. These results indicate that both models successfully

Index	IMDb ID	Movie Title	Genres	Subtitle Count
1	499448	The Chronicles of Narnia: Prince Caspian	Action, Adventure, Family, Fantasy	716
2	780521	The Princess and the Frog	Animation, Adventure, Comedy, Family, Fantasy, Musical, Romance	968
3	796366	Star Trek	Action, Adventure, Sci-Fi	640
4	800369	Thor	Action, Fantasy	892
5	810900	High School Musical 2	Comedy, Drama, Family, Music, Musical, Romance	1258
6	815178	The Life Before Her Eyes	Drama, Fantasy, Mystery, Thriller	410
7	864761	The Duchess	Biography, Drama, History, Romance, Thriller	880
8	892318	Letters to Juliet	Adventure, Comedy, Drama, Romance	427
9	989757	Dear John	Drama, Romance, War	820
10	4007502	Frozen Fever	Animation, Short, Adventure, Comedy, Family, Fantasy, Musical	66
11	413300	Spider-Man 3	Action, Adventure, Sci-Fi	787
12	421206	Gridiron Gang	Biography, Crime, Drama, Sport	916
13	443489	Dreamgirls	Drama, Music, Musical	1349
14	461770	Enchanted	Animation, Adventure, Comedy, Family, Fantasy, Musical, Romance	648

Table 9: List of Movies used in the Test dataset with IMDb ID, Title, Genres, and Subtitle Count.On average, each movie has a summary containing 601 words.

Index	IMDB ID	Movie Title	Genres	Subtitle Count
1	3634326	Tomorrowland	Action, Adventure, Drama, Romance, Sci-Fi, Thriller	236
2	3622592	Paper Towns	Adventure, Comedy, Drama, Mystery, Romance	1077
3	884328	The Mist	Horror, Sci-Fi, Thriller	770
4	475290	Hail, Caesar!	Comedy, Drama, Mystery	951
5	368933	The Princess Diaries 2: Royal Engagement	Comedy, Family, Romance	981
6	988045	Sherlock Holmes	Action, Adventure, Mystery	933
7	2334873	Blue Jasmine	Comedy, Drama, Romance	622
8	1854564	Percy Jackson: Sea of Monsters	Adventure, Family, Fantasy	992
9	213149	Pearl Harbor	Action, Drama, Romance, War	911
10	1924435	Let's Be Cops	Action, Comedy, Crime	1727
11	2379713	Spectre	Action, Adventure, Thriller	708
12	1905041	Fast & Furious 6	Action, Thriller	935
13	1837703	The Fifth Estate	Biography, Crime, Drama, Thriller	940
14	2398241	Smurfs: The Lost Village	Animation, Adventure, Comedy, Family, Fantasy	1070
15	1840309	Divergent	Action, Adventure, Mystery, Sci-Fi	783
16	2132285	The Bling Ring	Biography, Crime, Drama	401
17	404032	The Exorcism of Emily Rose	Drama, Horror, Thriller	872
18	330373	Harry Potter and the Goblet of Fire	Adventure, Family, Fantasy, Mystery	917
19	4846340	Hidden Figures	Biography, Drama, History	1358
20	800039	Forgetting Sarah Marshall	Comedy, Drama, Romance	1723

Table 10: List of Movies used in the Train dataset with IMDb ID, Title, Genres, and Subtitle Count.On average, each movie has a summary containing 700 words.

conveyed the figurative meaning of idiomatic expressions but often relied on literal or descriptive translations rather than direct idiomatic equivalents in Spanish. These findings suggest that, while the models capture the essential sense of the idioms, there remains room for improvement in achieving more culturally nuanced and idiomatically faithful translations.

The idiom "time will tell" conveys the idea that the outcome of a situation will become clear only after some time has passed. As shown in Table 14, both GPT-4o and LLaMA-3 translated this phrase as "Solo el tiempo lo dirá" across all prompts. This translation is a well-established equivalent in Spanish, accurately preserving both the figurative meaning and natural phrasing of the original expression.

However, for "I'm completely out of countenance" as shown in Table 15, GPT4o produced the expected idiomatic translation "Estoy completamente desconcertado" closely matching the reference and preserving the intended meaning. In contrast, LLaMA-3 generated varied outputs, such as "Estoy completamente fuera de lugar" (out of place) and "Estoy completamente fuera de mí" (beside myself). While these translations convey a related emotional state, they alter the nuance and do not fully retain the idiomatic meaning, highlighting inconsistencies in LLaMA-3's handling of idioms.

Colloquialisms In the Table 16 compares how GPT-4o and LLaMA-3 handle slang phrase in translation, using the phrase "pop the question" a casual way of saying "propose marriage." GPT-4o translates it as "hacer la gran pregunta," which re-

ID	En→Es			En→De			En→Fr			En→Fi		
	GPT-4o	GPT-3.5	LLaMA-3	GPT-4o	GPT-3.5	LLaMA-3	GPT-4o	GPT-3.5	LLaMA-3	GPT-4o	GPT-3.5	LLaMA-3
simple	59.28	54.07	56.25	53.87	47.21	51.47	52.92	49.58	49.58	45.15	45.15	45.73
movie domain	59.88	59.07	56.63	54.34	53.39	51.40	52.85	50.02	50.02	50.93	50.93	46.82
+ N = 2	60.07	50.93	56.86	54.70	53.08	51.34	53.09	51.98	50.37	53.20	50.39	46.47
+ N = 3	60.16	59.20	56.92	54.71	53.51	51.44	53.21	52.14	50.32	53.21	50.78	46.62
+ N = 4	60.21	59.31	56.93	54.63	53.54	51.53	53.16	52.23	50.39	53.40	50.97	46.64
+ N = 5	60.11	59.28	56.88	54.79	53.47	51.49	53.23	52.19	50.43	53.29	50.90	46.53
+ title	60.13	59.28	**56.72**	**54.77**	**53.60**	**51.63**	**53.22**	**50.24**	**50.24**	**50.97**	**50.97**	**46.86**
+ summary	60.15	**60.33**	55.50	54.66	53.37	50.78	53.15	49.90	49.90	50.86	50.86	46.77
+ genre	60.04	59.13	56.58	54.53	53.43	51.08	52.94	50.07	50.07	50.83	50.83	46.75
all	**60.24**	58.58	55.71	54.65	53.38	50.95	53.11	49.73	49.73	50.86	50.86	46.72
title + N = 4	60.23	59.39	57.01	54.91	53.71	51.62	53.41	52.36	50.48	53.48	51.09	46.62

Table 11: chrf++ for prompts including meta-information and previous context for GPT-3.5, GPT-4o, and LLaMA-3 models. The rows labeled N=2 to N=5 show the results of using previous context lines in the prompt. The highest scores for meta-information are in bold, while the highest scores for context are underlined. Cells highlighted in red indicate the overall highest scores across both meta-information and context.

ID	En→Es		En→De		En→Fr		En→Fi	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
M2M	21.7	0.7902	18	0.7502	17.4	0.7906	11.8	0.7906
simple	31.36	0.8175	23.57	0.7862	24.65	0.7444	14.35	0.8186
movie domain	31.96	0.8344	24.24	0.8028	25.11	0.7861	15.39	0.8238
+ N = 2	32.05	0.8366	24.17	0.8009	24.70	0.7796	15.09	0.8307
+ N = 3	32.10	0.8368	24.24	0.8013	24.59	0.7783	15.12	0.8207
+ N = 4	32.11	0.8369	24.37	0.8013	24.60	0.7793	15.17	0.8208
+ N = 5	32.06	0.8365	24.26	0.8011	24.61	0.7792	15.07	0.8201
+ title	**32.15**	**0.8413**	**24.42**	**0.8117**	**25.33**	**0.7946**	**15.59**	0.8084
+ summary	32.04	0.8136	24.37	0.7607	25.14	0.7660	15.50	0.8031
+ genre	31.94	0.8163	24.15	0.6644	25.07	0.7669	15.38	0.8063
+ all	32.05	0.8144	24.41	0.7829	25.12	0.7650	15.44	0.8028
title + N = 4	32.08	0.8376	24.34	0.8107	26.02	0.7902	15.02	0.8309

Table 12: COMET and BLEU scores for zero-shot prompts including meta-information and previous context for GPT-3.5, GPT-4o, and LLaMA-3 models. The rows labeled N=2 to N=5 show the results of using previous context lines in the prompt. The highest scores for meta-information are in bold, while the highest scores for context are underlined. Cells highlighted in red indicate the overall highest scores across both meta-information and context with new additional data.

```
{The following is taken from the subtitles of the
movie {title}. Translate it from English to [tgt]
English: [en_sentence]
[tgt]: [tgt_sentence] }x K

The following is taken from the subtitles of the movie
{title}. Translate it from English to [tgt]
English: [en_sentence]
[tgt]:
```

Table 13: Prompts used in K-shot learning. The substrings within are repeated K times.K= 0, 3, 5

English:	`Only time will tell`
Spanish:	`Solo El tiempo lo dirá`
GPT-4o	
all the prompts	`Solo el tiempo lo dirá`
LLaMA-3	
all the prompts	`Solo el tiempo lo dirá`

Table 14: Example of a translation from English to Spanish, including an idiomatic expression, generated by GPT-4o and LLaMA-3.

tains the expressive and conversational tone, while LLaMA-3 translates it as "hacer la pregunta" a more neutral version that loses some of the original informal style.

Table 17 examines how both models translate colloquial speech in "That'll go down better with white folks". "White folks" is a colloquial and informal way of referring to white people, commonly used in conversational English, particularly in American English, and often carries a regional, cultural, or social nuance, depending on the context. While "los blancos" aligns more closely with the informal tone of the original phrase, "la gente blanca" softens the expression, making it sound more neutral and potentially more appropriate in formal contexts. These examples show how GPT4o tends to preserve slang and informal expressions more naturally, while LLaMA-3 often produces a more literal or neutral translation, sometimes softening colloquial terms.

Table 18 shows that although both models correctly translated "wee bit" as "un peu", but misinterpreted "dodgy", which in this case referred to a machine being worn out or rusty (rouillée) rather than suspicious. GPT-4o translated it as "douteuse" (doubtful), while LLaMA-3 rendered it as "louche"

English:	I’m completely out of countenance
Spanish:	Estoy absolutamente desconcertado
	GPT-4o
all the prompts	Estoy completamente desconcertado
	LLaMA-3
simple	Estoy completamente fuera de lugar
movie domain title summary genre all	Estoy completamente fuera de mí

Table 15: Example of a translation from English to Spanish, including an idiomatic expression, generated by GPT-4o and LLaMA-3.

English	Oh, so you want to pop the question tonight, huh?
Spanish	oh, Así que esta noche quiere pedirle la mano, ¿eh?
	GPT-4o
title	Oh, ¿así que quieres hacer la gran pregunta esta noche, eh?
	LLaMA-3
title	¡Ah, así que quieres hacer la pregunta esta noche, eh?

Table 16: Example of a translation from English to Spanish, including slang, generated by GPT-4o and LLaMA-3.

(suspicious), highlighting the challenge of accurately translating slang and colloquial expressions related to mechanical conditions without explicit clarification.

D Models Sensitivity to Prompts

Across the board, GPT-4o consistently outperformed GPT-3.5 and LLaMA-3.

Several key issues were observed. In many cases, models such as GPT-4o and LLaMA-3 copied source phrases or parts of the prompt template itself with the target translation as in Table 19.

LLaMA-3, while generally lagging behind the other models, shows a decline in performance when dealing with prompts that include extensive meta-information, such as summaries. When summaries are included in the prompts, the model sometimes struggles to produce a coherent translation when some words are not in the provided in summary. Example is given in the Table 22. When analyzing the output, LLaMA-3 often provides additional information, such as explanations, unless system instructions are explicitly given to provide only the translation, as seen in Table 21. Although these explanations may not be necessary for the translation task, they indicate that the model is attempting to convey an understanding of the context, such as recognizing the informal nature of subtitles, which are often presented in colloquial language.

English:	That’ll go down better with white folks.
Spanish:	A los blancos les va a gustar más.
	GPT-4o
title	Eso caerá mejor con los blancos
	LLaMA-3
title	Eso caerá mejor con la gente blanca

Table 17: Example of a translation from English to Spanish, including slang, generated by GPT-4o and LLaMA-3.

English:	Well, she’s a wee bit dodgy
French:	Eh bien, elle est un peu rouillée
	GPT-4o
title	Eh bien, elle est un peu douteuse
	LLaMA-3
title	Eh bien, elle est un peu louche

Table 18: Example of a translation from English to French, including an idiom and slang, generated by GPT-4o and LLaMA-3.

E License

In our study, we utilized the OpenSubtitles dataset, which was available under the GNU General Public License v3.0 (GPLv3). Wikipedia summaries were used in accordance with the Creative Commons Attribution-ShareAlike 3.0 (CC BY-SA 3.0) license, with proper attribution and shared derivative works. IMDb data was accessed for non-commercial research purposes, following IMDb’s Terms of Use with appropriate attribution.

The following is taken from the subtitles of the movie "Thor". Translate it from English to Spanish
English:Thank you for your cooperation.
Spanish:

Answer
In Spanish, “Thank you for your cooperation.” translates to:“Gracias por su cooperación.”

Table 19: LLM’s response copied source phrases or parts of the prompt template.

English: What is the formula for the volume of a sphere?
[German]:

Answer
German: Die Formel für das Volumen einer Kugel lautet:
[V = $\frac{4}{3}\pi r^3$]
Dabei ist V das Volumen und r der Radius der Kugel.

Table 20: GPT-4o’s tendency to output answers to questions instead of providing a direct translation when the source sentence is a in Question format when using the*simple*.

The following is taken from the subtitles of the movie Spider-Man 3.Translate it from English to German.
English: Just had bad luck
[German]:

Answer
German: Hatte einfach Pech gehabt.
(Note: **"Pech" is a colloquialism** in German that roughly translates to “bad luck” or “misfortune", and is often used in informal settings.

The following is taken from the subtitles of a movie. Translate it from English to Spanish.
English: That’s got some zang to it!
[Spanish]:

Answer
¡Eso tiene un poco de salsa!.
“Zang" is a colloquialism used to describe something with a lot of flavor or energy, so I translated it to “un poco de salsa", which means “a little bit of flavor" or "some zest"

Table 21: Examples of translation outputs generated by LLaMA-3 for English to German and English to Spanish translations. The examples highlight the model’s ability to understand colloquial expressions.

Here is a summary of a movie: summary. The following is taken from the subtitles of that movie. Translate it from English to Finnish.
English: The dog just spoke to me!"
[Finnish]:

Answer
There is no mention of a dog speaking to anyone in the provided movie summary. If you meant to provide a different quote from the movie, please feel free to share it, and I’ll be happy to help you translate it into Finnish.

Table 22: Example where LLaMA-3 fails to provide a translation in a prompt containing a summary.