Leveraging Domain Knowledge at Inference Time for LLM Translation: Retrieval versus Generation

Bryan Li[*]
University of Pennsylvania
bryanli@seas.upenn.edu

Jiaming Luo, Eleftheria Briakou, Colin Cherry
Google
{jmluo,ebriakou,colincherry}@google.com

Abstract

While large language models (LLMs) have been increasingly adopted for machine translation (MT), their performance for specialist domains such as medicine and law remains an open challenge. Prior work has shown that LLMs can be *domain-adapted* at test-time by *retrieving* targeted few-shot demonstrations or terminologies for inclusion in the prompt. Meanwhile, for *general-purpose* LLM MT, recent studies have found some success in *generating* similarly useful domain knowledge from an LLM itself, prior to translation. Our work studies domain-adapted MT with LLMs through a careful prompting setup, finding that demonstrations consistently outperform terminology, and retrieval consistently outperforms generation. We find that generating demonstrations with weaker models can close the gap with larger model's zero-shot performance. Given the effectiveness of demonstrations, we perform detailed analyses to understand their value. We find that domain-specificity is particularly important, and that the popular multi-domain benchmark is testing adaptation to a particular writing style more so than to a specific domain.

1 Introduction

Large language models (LLMs) have emerged as the next major paradigm for machine translation (MT), with increasing use in both industrial and academic settings. These models are exciting not only for their strong base (or *zero-shot*) translation capabilities, but also for their ability to be modified at inference time through alternate prompts (Kojima et al., 2022; Kong et al., 2024), in-context learning (Brown et al., 2020) and the use of intermediate reasoning (Wei et al., 2024).

This flexibility is particularly exciting for adapting LLMs to translate specialist domains, such as legal or medical texts. In the statistical and neural MT eras, domain adaptation would typically take the form of an expensive continued training procedure on in-domain data (Freitag and Al-Onaizan, 2016; Thompson et al., 2019). With LLMs, there is the promise of simple adaptation at inference time.

One promising technique is the retrieval of instance-specific demonstrations of translation from a bitext datastore for few-shot in-context learning, which has shown large improvements for domain-adapted MT (Agrawal et al., 2023; Tan et al., 2024), rivaling the performance of specialized nearest-neighbor MT systems (Khandelwal et al., 2021). LLMs have also been shown to make good use of bilingual terminology dictionaries for lexical translation hints (Ghazvininejad et al., 2023; Lu et al., 2023; Moslem et al., 2023).

Intriguingly, two recent approaches have forgone external resources in favor of querying an LLM to generate useful knowledge from its internal memory. First, the MAPS approach issues LLM queries for topics, terminology, and demonstrations based on the source text (He et al., 2024). Their terminology and demonstrations mirror the knowledge sourced from retrieval steps in earlier work. The idea is that the LLM has seen relevant information during pre-training, and would benefit from explicitly surfacing it before translation. Second, the step-by-step MT approach queries its LLM to translate and discuss idiomatic phrases before performing a complete translation (Briakou et al., 2024). However, both these works only consider the general domain. This inspires us to consider the applicability of internal memory approaches to domain adaptation, for which relevant external resources may be more difficult to obtain.

In this work, we study the effectiveness of different representations of domain-specific knowledge, in *strategies* – external retrieval vs. internal generation – and *sources* – translation demonstrations and bilingual terminology. We consider three domains (law, medical, and Koran) from the commonly-

[*]Work done at an internship at Google Translate Research.

Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing (KnowledgeNLP'25), pages 91–106

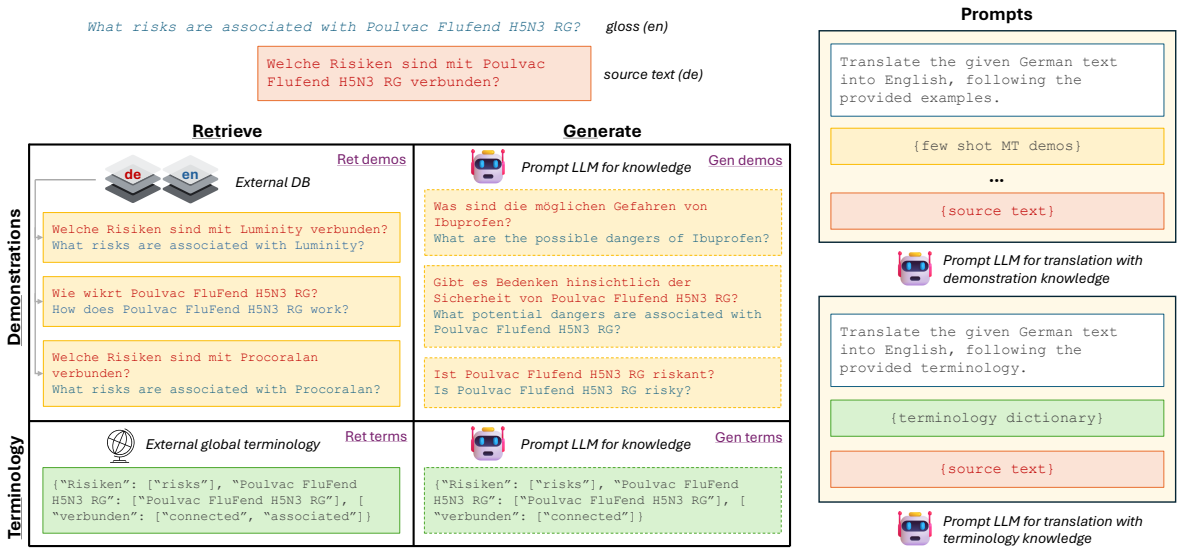

Figure 1: Illustration of the main MT settings, for an example source text in German. The two knowledge strategies are demonstrations vs. terminology; the two sources are retrieval vs. generation. This gives 4 settings for comparison. Within a strategy, we use the same prompts, varying only the provided information.

used multi-domain dataset (Aharoni and Goldberg, 2020), and experiment with two LLMs (Gemini 1.5 and Gemma-2). Our study addresses three main research questions:

RQ1. For improving domain-adapted MT, how viable is generation from an LLM's parametric memory compared to retrieval from external resources?

RQ2. Likewise, how does adapting MT with demonstrations compare with terminologies, regardless of their source?

RQ3. Given the effectiveness of demonstrations, can we attribute which of their aspects contribute the most for both retrieval and generation?

We discuss knowledge sourced from retrieval in §2.1 and from generation in §2.2. Comparisons between terminology and demonstrations are enabled by our use of a silver terminology dictionary, built by LLM analysis of the same bitext used as the datastore of demonstrations (§3.1). This allows us to study demonstrations and terminology as alternate views into the same base data in the retrieval setting. We address RQ1 and RQ2 with the results in §4. We explore RQ3 through several analyses in §5; the main takeaways are that retrieved demonstrations mainly provide hints of target style rather than terminology, and that generated ones can viably boost performance, albeit to the same level as static domain-specific demonstrations.

2 Leveraging Domain Knowledge

Comparisons between representations of domain knowledge are enabled by our careful prompting setup which decouples the *source* and *strategy*, as sketched in Figure 1. Bilingual terms, whether retrieved externally or generated by an auxiliary LLM call, feed into the same translate-with-terms prompt, and likewise for demonstrations. On *sources*, retrieval leverages resources such as datastores and dictionaries, while generation elicits information from an LLM's own parametric memory. On *strategies*, demonstrations provide source-target example pairs, whereas terminology focuses on domain-specific lexical items. This section details the integration of these strategies and knowledge sources within our experimental framework for domain adaptation of LLM MT.

2.1 Knowledge from Retrieval

We describe two successful approaches to retrieve domain knowledge from external resources: *demonstration retrieval* and *terminology lookup*. The two related approaches operate in different fashions. Demonstration retrieval has the model *implicitly* learn from the characteristics of the exemplars, both style and terminology. Terminology lookup has the model *explicitly* see which source terms are important and also how to translate them.

Resource requirements These methods, while effective, are expensive, as they require the ex-

istence of high-quality and domain-specific resources. The former requires a large pool of bitext demonstrations, and the latter requires the creation of a term-rich bilingual dictionary.

2.1.1 Demonstration Retrieval

Demonstrations are provided as exemplars in the prompt to facilitate in-context learning (ICL) (Brown et al., 2020; Patel et al., 2023). These exemplars can either be *static*, the same across all instances, or *instance-specific*, in which different exemplars are retrieved for each instance to provide specific guidance and hints.

The typical setup for demonstration retrieval for MT is as follows. Given a source text, we find k closest source-side matches in an external datastore, using some similarity metric, such as BM25 or cosine similarity of embedding vectors. Then, we include in the LLM prompt these k source texts paired with the gold target translations.

Prior work The use of demonstrations has a long history in MT, with some of the oldest data-driven approaches to MT having as their first step finding the most relevant examples from a bilingual translation memory. This idea has been used for computer-aided translation (Yamada, 2011), example-based MT (Somers, 1999; Lepage and Denoual, 2005) and statistical MT (Koehn and Senellart, 2010).

Several recent papers have studied what constitutes effective demonstration retrieval for MT with LLMs, with a particular focus on the multi-domain dataset. Agrawal et al. (2023) found a strong baseline to be example-specific BM25 retrieval of bitexts, which can be strengthened further by re-ranking for lexical diversity. Tan et al. (2024) use a much larger LLM, and show that BM25 retrieval of target sentences alone can compare favorably with both sides of bitexts. Conversely, in the general-domain, researchers have found that a demonstration's quality matters more than its proximity (Vilar et al., 2023; Zhang et al., 2023).

Our Setup For our few-shot implementation, we design a simple prompt (shown in Figure 5). We use k=3 exemplars, [1] and retrieve using the BM25 metric. Our datastore, derived from the train split of multi-domain, has 16,775 demonstrations for Koran, 234,352 for medical, and 464,295 for law.

[1]Prior work often chooses $k \geq 10$. As we find COMET for k=3 and k=10 differ by ~0.3 we thus choose k=3 to fairly compare to the 3 generated demonstrations in a later setting.

2.1.2 Terminology Lookup

Intuitively, one of the major challenges when translating in a specialist domain is the adaptation to domain-specific terminology. Especially in high-stakes legal, medical or business domains, precision of terminology can be crucial. Bilingual dictionaries of terminology are therefore likely sources of useful external knowledge to add into an MT system. These resources can be easier to construct than the large datastore of translation demonstrations needed in §2.1.1. In fact, the construction of a clear terminology may very well be a prerequisite to creating human translations.

Prior Work Improving translations with terminologies has been heavily studied. In the statistical and neural eras, solutions could take the form of incorporating dictionaries into training (Wu et al., 2008), or controllable MT systems that respect example-specific terminology constraints included in the input (Post and Vilar, 2018; Wang et al., 2022). More recently, terminology constraints have been studied at two WMT shared tasks (Alam et al., 2021; Semenov et al., 2023). These approaches illustrate two different motivations for the use of terminology dictionaries in MT: the dictionary can be viewed as a useful source of domain-specific information, or as a set of constraints that must be followed consistently. Our work aligns with the former motivation, viewing bilingual terminologies only as hints to improve overall quality.

With the advent of LLMs, terminologies can be included in the prompt, with additional instructions on their usage. Most LLMs follow these instructions easily, as shown at the WMT23 shared task on terminology (Semenov et al., 2023). For example, Moslem et al. (2023) find that for the COVID-19 domain, a prompt using retrieved terminologies significantly boosts term success rate and also improves human evaluation scores. Other works have explored how to more effectively format dictionaries (Lu et al., 2023; Ghazvininejad et al., 2023).

Our Setup Since the multi-domain dataset does not have a provided domain-specific terminology, we derive one from the multi-domain training set, as described in §3.1. Keeping with our theme of providing hints rather than constraints, the dictionary gives a list of possible translations for each source term, each licensed by at least one example in the training set.

With this dictionary in place, we look up terms

by exact lexical match to the source text currently being translated, and include any matches in our prompt for translation with terminology (shown in Figure 6). The LLM is instructed to pick the most appropriate translation among the choices, given the source. Note that the translation prompt also includes three domain-specific examples of how to translate with terminologies.

2.2 Knowledge from Generation

While external knowledge retrieval demonstrably benefits knowledge-intensive NLP tasks, whether it is truly necessary for domain-adapted MT still warrants investigation, given that LLMs are explicitly trained on massive corpora including texts from specialist domains. Therefore, we investigate whether leveraging LLMs' internal parametric memory can offer comparable benefits, and thus circumvent the costly acquisition and curation of external resources. This approach effectively simulates external retrieval by prompting the LLM to generate relevant information.

Resource requirements By design, the generation setting requires almost no external resources. The approaches discussed below only required us to manually create a handful of static exemplars for each subtask, which are used for all of its prompts.

2.2.1 Prior Work

Prior work has explored several methods to leverage an LLM's parametric knowledge to improve MT quality, either post-translation, or pre-translation. Most relevant to our work are two studies which operate at the pre-translation stage.

He et al. (2024) propose a human-like translation process, where they separately prompt LLMs for 3 aspects related to a source text (demonstrations, topics, and terms). Directly using these generated knowledge pieces in another LLM interaction is insufficient, and so they rely on an external quality estimation (QE) method to select among candidates, improving general domain MT quality. Our generation setting also use demonstrations and terms, but without any external feedback from QE.

Briakou et al. (2024) propose a method to model the LLM translation process step-by-step. Their 2-step approach has an LLM first perform research on idiomatic expressions, then perform the full translation. For document-level MT datasets, they find this consistently outperforms zero-shot MT.

2.2.2 Demonstration Generation

We author a prompt to generate demonstrations (Figure 8). For each domain, we provide 3 example demonstrations for 2 static, real source sentences. This is inspired by the demonstration aspect of He et al. (2024), but we elicit 3 demonstration pairs at a time instead of 1.

Best practices To easily parse the 3 demonstration pairs, we ask for a prescribed JSON output format. We also find that providing static few-shot exemplars of the demonstration task is key to both diversity among the 3 demonstrations, and output format adherence. We use a different set of exemplars for each domain, drawn from the train set. We perform ablations on the contributions of different aspects of generated demonstrations in §5.2.

2.2.3 Terminology Generation

We design a prompt to generate terminologies from a single source sentence (Figure 10), also using 2 static, real sentence pairs for each domain. This follows in the spirit of the research step of Briakou et al. (2024), where they explain this as having the LLM perform intermediate reasoning about hard-to-translate parts. However, there are several differences resulting from their focus on document-level MT. We ask generally for terminologies, while they ask specifically for idiomatic expressions, which are more prevalent in long documents. We also prescribe a JSON format (same as for retrieved terms), while theirs allows for free-form output.

Best practices We again found that best performance is achieved with static, domain-specific few-shot exemplars of the terminology task, and the prescribed JSON format.

3 Experimental Setup

Dataset We experiment with the multi-domain dataset (Aharoni and Goldberg, 2020), using the filtered version provided by Tan et al. (2024), with 3 domains: law, medical, and Koran. Multi-domain covers the German-English (de-en) direction, and consists of dev and test sets, with ~2000 entries per domain, as well as a train set with 1M+ entries.

LLMs We perform experiments with two LLMs, the open LLM Gemma-2 27B IT (Team, 2024b), and the proprietary Gemini 1.5 Pro (Team, 2024a). We thus can investigate which settings, if any, are more effective with the smaller model vs. a much larger model respectively.

Evaluation We perform zero-shot MT as a baseline, and employ the four settings described in §2 for comparison: retrieved demonstrations, retrieved terminologies, generated demonstrations, and generated terminologies. Appendix B lists all prompts used in this work. Following Vilar et al. (2023), we use a neural automated metric, COMET (Rei et al., 2022). While prior work also considered the lexical metric BLEU, we found that it was overly sensitive to minor rephrasing. This is in line with studies that show neural metrics correlate much better with human judgments of LLM translation quality (Freitag et al., 2021; Kocmi et al., 2021).

3.1 Terminology Dictionary Creation

Our multi-domain test scenario does not come with bilingual terminology dictionaries for its domains. However, we can create them from the provided training split, following the methodology in prior work (Moslem et al., 2023; Semenov et al., 2023).[2] We design a prompt (Figure 12) to extract terminologies from a given source-target text pair, providing 5 static exemplars to demonstrate what is meant by "terminology". We then apply this to each pair from the train split. Then, we aggregate all of the output terms, to get one large dictionary with one-to-many mappings.[3] We create a separate global terminology for each of the three domains.

Given the large size of the training split (700K entries), we make two adjustments to reduce the number of model calls. First, we batch five test pairs at a time into a single call. Second, we consider only the subset of train entries that were ever retrieved by BM25 over the test set (i.e. the entries that are actually relevant); this constitutes 70K entries, or about 10% of the total entries.

Note that the train split is also used for demonstration retrieval, therefore enabling a controlled comparison between the two external knowledge sources. Furthermore, unlike prior work using one-to-one terminology mappings, we explore a more realistic one-to-many scenario, with all possible translations in the prompt for the LLM to select.

4 Results

Table 1 presents our primary results, comparing LLM translation enhanced with domain-specific knowledge in the form of translation demonstrations or bilingual terminology, with the artifacts derived from either external retrieval (§2.1) or internal generation (§2.2). First, in line with prior work, we confirm that retrieved demonstrations improve over zero-shot across models and domains studied. We next describe the three main findings.

Demonstrations outperform terminology For all models and domains studied, knowledge provided in the form of demonstrations consistently outperforms terminology. For Gemma, we see that all settings improve performance,[4] but the improvements from demonstrations are markedly larger. The differential is more pronounced for Gemini, which starts from a much stronger baseline than Gemma. Terms, either retrieved or generated, do not provide much of a boost over zero-shot for Gemini, while demonstrations result in significant improvements. The takeaway for this finding is that for weaker models, providing domain knowledge from any source or strategy is beneficial. Conversely, stronger models do not benefit from domain-specific terminology, but only from more complete demonstrations of the task.

Retrieval outperforms generation The second notable trend across models and domains is that retrieval consistently outperforms generation. With Gemma, demonstration generation outperforms zero-shot by +2.3 (averaged across domains), while retrieval further improves to +3.4. For the more powerful Gemini, the differential is larger – demonstration generation outperforms zero-shot by +0.5, while retrieval by +1.8.

Generated domain-specific demonstrations boost weaker model's translations Taking the prior two findings together, we can bootstrap domain-adapted MT knowledge from an LLM's own parametric memory, with the two-stage approach of first generating demonstrations, then translating. This improvement especially pronounced with Gemma (+2.3 vs. +0.5 over respective zero-shot). In fact, this empowers a smaller model (Gemma) to close the gap with a larger model's (Gemini) zero-shot results, as can be seen by comparing, in Table 1, the bottom left and top right rows. The gains in medical (+2.9) and Koran domains (+1.0) result in statistically

[2] We did not perform human post-editing due to the datastore's size (700K), but we note in an experiment by Moslem et al. (2023), they found humans rated 95%+ terms as accurate.

[3] For quality controls, we kept only entries where 1) target terms have >10% usage and 2) both sides of terms match.

[4] To explain the outliers for Koran (-1.4, -0.5), our manual analysis found term inconsistency – high-frequency source terms mapped to multiple, equally-valid target terms.

Domain Knowledge?		Gemma-2 27B IT						Gemini 1.5 Pro					
		Law		Med.		Koran		Law		Med.		Koran	
zero-shot		84.8		85.2		75.1		86.6		88.2		76.3	
retrieved	terms	85.9*	↑1.1	87.8*	↑2.6	74.6	↓0.5	86.9*	↑0.3	88.5	↑0.3	74.9	↓1.4
	demos	88.6*	↑3.8	89.9*	↑4.7	76.7*	↑1.6	89.3*	↑2.7	89.9*	↑1.7	76.4	↑0.1
generated	terms	85.2	↑0.4	87.1*	↑1.9	75.7*	↑0.6	86.7	↑0.1	88.1	↓0.1	76.9*	↑0.6
	demos	86.0*	↑1.2	88.1*	↑2.9	76.1*	↑1.0	87.2*	↑0.6	88.8*	↑0.6	76.7*	↑0.4

Table 1: Results for MT using the COMET22 metrics, comparing the knowledge sources, retrieved and generated, and the strategies, demonstrations (demos) or terminology (terms). Significant improvements ($p < 0.05$) over the zero-shot baseline are marked with *. Demonstrations outperform terminology, and retrieval outperforms generation. Generation is especially effective for the smaller Gemma model.

equivalent scores. Law domain incurs a decent gain (+1.2), but still is below Gemini (86.0 < 86.6). It is worth emphasizing that acquiring extensive resources for novel specialist domains is expensive; but this straightforward approach can be effective.[5]

Comparisons with Prior Results We can also compare our results with demonstrations to those from the recent study by Tan et al. (2024), who use the `gpt-3.5-turbo-0301` LLM. Their zero-shot results are most comparable to Gemma's: 84.4, 86.2, 75.1. Their results for retrieved demonstrations are also comparable: 88.2, 89.6, 76.5. The other 3 settings, retrieving terms and both generation ones, are new to our work – and we re-emphasize here the value of our controlled setting in facilitating fair comparison between them all.

5 Analysis

Demonstrations (both retrieved and generated) are by far the most effective domain adaptation strategy we explored, providing a large boost to both LLMs. In the following sections, we turn to analyses to understand better where the gains are coming from. We begin by analyzing retrieved demonstrations to disentangle contributions from style vs. terminology (§5.1). Then, we investigate the importance of various in-context learning decisions for generating demonstrations (§5.2). Finally, we study how generated domain knowledge can be distilled at test-time from larger to smaller models (§5.3).

5.1 Retrieved Demonstrations: Contributions from Style vs. Terminology

What exactly is being conveyed by the retrieved demonstrations? In this section, we take advantage of our careful experimental setup, where our

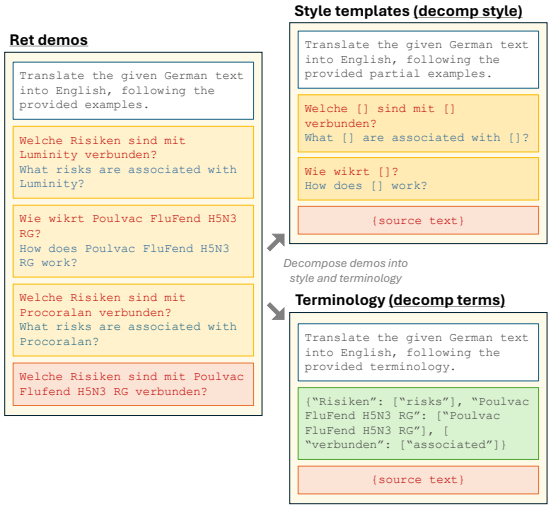

Figure 2: Illustration of our process to decompose the contributions of retrieved demonstrations into style and terminology. We first extract the source-target term pairs using a simple function, and aggregate them into a local terminology. Then, the remaining tokens are the style templates, with the terms masked. Note that in the actual data, we use <MASK> instead of [].

bilingual terminology is derived from the same parallel text used for demonstrations, to disentangle whether demonstrations are more valuable because they assist with proper *terminology* translations in context, or with matching the *style* of the corpus.

The core idea behind this experiment is that we can use the same technique to extract bilingual terminology pairs from a translation demonstration (§3.1), but instead of running it on the whole training corpus, we can run it only on the k demonstrations retrieved to match the current source sentence. This gives us a *local* terminology, as opposed to a *global* one. Crucially, where the global terminology would present the union of all possible target language translations found throughout the training set for a given source term, the local terminology only presents translations licensed by the k demonstrations. This allows it to take advantage of any

[5]Note that the experiment from §5.2 shows that domain-specificity is the main contributor, rather than proximity to the current instance. These generated demos are only as effective as real static, but domain-specific demos.

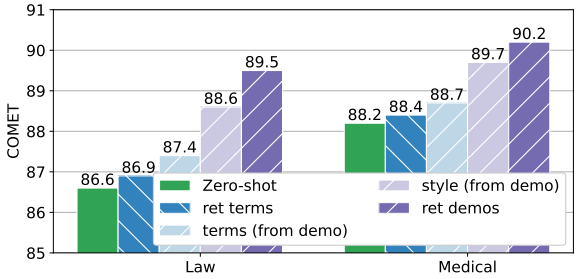

Figure 3: Results for zero-shot, external retrieval, terms from demonstrations, and style from demonstrations.

disambiguating context in the demonstrations to create more relevant term translations.

We then define style templates as the inverse – the remaining tokens, with the bilingual terms on both sides replaced with a <MASK> mask token. For this, we use a similar prompt as for demonstration retrieval, but also explicitly instruct the LLM to not generate mask tokens in its output (as shown in Figure 7). Upon manual inspection, these masks appear quite thorough, with most anything that could be considered terminology being masked out.

Results We carry out the decomposition experiment using Gemini 1.5 Pro. Figure 3 presents our results. We see that compared to zero-shot, using local terms (terms from demonstrations) more than doubles the gains of global terms (retrieved terms). However, style templates (style from demonstrations) further narrow the gap to retrieved demonstrations by 60% (law) and 75% (medical).

The combined results from the *terms from demonstrations* and *style from demonstrations* experiments indicate that the primary value from retrieved demonstrations is not contextually appropriate translations of domain-specific terminology. While this is a part of the story, it accounts for only 0.8 (law) and 0.5 (medical) points of the 2.9- and 2.0-point improvements from the retrieved demonstrations. Meanwhile, the *style from demonstrations* scores almost perfectly account for the remainder. This is a strong indicator that the majority of the value of retrieved demonstrations comes from matching the publication style of these corpora, rather than carrying out adaptation to a medical or legal domain. That is, we are doing domain adaptation, but it is to a much more narrow domain than is usually discussed.

These results agree with and reinforce conclusions from recent work. Tan et al. (2024) perform a targeted study into translation style, following the same settings – the multi-domain dataset and a strong proprietary LLM. Their findings between zero-shot and few-shot concur – while there is a observable COMET difference (2.7), nevertheless zero-shot translations "have already conveyed the *semantic meaning* of the source sentence, albeit with some variations in *lexical choices* and sentence structure." They therefore propose a style learning method to retrieves related target sentences from a monolingual target corpus, finding this achieves 70% of few-shot's gains. However, by only removing the source side of the demonstrations, the exemplars still implicitly provide both style and terminology hints. We add to the discussion by providing a precise, alternative definition of "style" as anything outside of terminology. This in turn allows us to cleanly decompose the tokens from each demonstration into two subsets, and assign credit accordingly.

5.2 Ablation on Generated Demonstrations

Generation of demonstrations (`generate demos`) is by far the most successful of the two approaches. As described in §2.2, we made several decisions here: 1) using domain-specific exemplars; 2) using the intermediate generation of demonstration step; 3) in that step, selection of the ICL exemplars. We explore the impact of the decisions by comparing the `zero-shot`, `retrieved demos`, and `generate demos` results to the following ablations:

Static few-shot Drop the generate demo step, and use the 2x3 domain-specific examples[6] directly as static demonstrations of translation. This investigates the impact of domain-specificity alone.

No ICL With the generated demo step, but remove all exemplars from that step's instructions. This investigates the impact of ICL at all.

General ICL With the generated demo step, but use the 5x1 general-domain examples from He et al. (2024) instead of the 2x3. This investigates the impact of the domain-specificity of ICL.

Results are shown in Table 2. First, we consider ablation results on Gemini. There is a large drop between `zero-shot` and 'no ICL' (e.g., 88.2 -> 83.8 for medicine). Our manual analysis of a few 'no ICL' outputs finds that the generated demonstrations on the target side are often quite lexically close; we hypothesize these are unhelpful and affect downstream translations. Comparing general ICL to `generated demos`, we see that roughly

[6] 2x3 means there are two example source texts, which are each followed by three example translation pairs.

Setting	Domain-specific	Generate demo step	Demo # ICL	Gemma-2 27B IT law	Gemma-2 27B IT med.	Gemma-2 27B IT Koran	Gemini 1.5 Pro law	Gemini 1.5 Pro med.	Gemini 1.5 Pro Koran
`zero-shot`	N/A	✗	N/A	84.8	85.2	75.1	86.6	88.2	76.3
static few-shot	✓	✗	N/A	86.3	88.2	76.3	87.2	89.1	76.4
`retrieved demos`	✓	✗	N/A	88.6	89.9	76.7	89.3	89.9	76.4
no ICL	N/A	✓	0	85.2	87.5	75.1	83.9	83.8	75.6
general ICL	✗	✓	5x1	85.7	87.8	75.6	86.9	88.5	76.0
`generated demos`	✓	✓	2x3	86.0	88.1	76.1	87.2	88.8	76.7

Table 2: COMET22 Results for the study on demonstration generation, using Gemma (left) and Gemini (right). The *italicized* settings are ablations, while the `monospace` settings are the same as in Table 1.

half the value of demonstration generation can be retained with general ICL. However, comparing 'static few-shot' to generate demos (rows 7 & 10), both achieve similar scores across domains. This adds a caveat to our earlier findings, suggesting the domain-specificity of the generated demos is more important to downstream MT than the demos alone.

Now, we consider ablation results on Gemma. Interestingly, unlike for Gemini, for Gemma even the 'no ICL' setting improves upon zero-shot (85.2 -> 87.5 for medicine). We observe that, compared to `generated demos`, 'General ICL' slightly underperforms it, while 'static few-shot' matches it. This again underscores the value of the demonstration stage in improving the smaller LLM's translations, as well as key role of domain-specificity.

Our results add insight into two formerly disparate findings. Prior work on older LLMs discussed two factors for ICL exemplars: lexical coverage within a domain (Agrawal et al., 2023), and their quality (Vilar et al., 2023). Our finding here provides evidence that, for current LLMs with strong zero-shot MT performance, the primary value of ICL is in the domain-specificity, especially in style. Quality examples can be equally as validly obtained from static few-shot exemplars or generated demonstrations.

5.3 Cross-LLM Knowledge Generation

For the two generation-based settings, the same LLM is used in both the generation stage and the translation stage. To further understand how generation quality affects the final performance, we conduct additional experiments to reuse the generated demonstrations or terminology from Gemini 1.5 Pro to prompt the Gemma 2 27B model for translation. As shown in Table 3, demonstration generation and terminology generation both benefit greatly from higher quality generations from Gemini, with significant gains in all three domains. This shows that higher-quality generated knowledge result in higher-quality translations. The larger Gemini model's knowledge can be effectively distilled to the smaller Gemma model, at inference-time, through its translation demonstrations.

strategy	gen. LLM	law	med.	koran
⚙ `demos`	Gemma	86.0	88.1	76.1
	Gemini	86.9*	88.6*	76.6*
⚙ `terms`	Gemma	85.2	87.1	75.7
	Gemini	85.8*	87.5*	76.4*

Table 3: Results for the ablation on generation-based strategies. Gemma-2 27B IT is always used for translation, but the generation model can be either LLM. Significant improvements when using Gemini's generated outputs instead of Gemma's are marked with *.

6 Discussion and Conclusion

We study the problem of domain adaptation for MT with LLMs, one which intuitively speaking, should be well addressed by prompting-time adaptation. Building upon prior work which injects domain-specific knowledge into prompts, we perform a thorough study into how this knowledge can best be acquired in terms of strategy, demonstrations or terminologies, and sources, retrieval or generation.

Our main study shows that demonstrations outperform terminology, and knowledge retrieval consistently outperforms generation. Furthermore, generation of domain-specific demonstrations can viably improve weaker model's performance, closing the gap with a larger model's zero-shot performance (though comparable to static exemplars). We gain additional insights with our further analyses. Notably, we explore the connection between the strategies, characterizing demonstrations as providing both terminology hints and style hints. Our

decomposition of the contributions of demonstrations finds that the majority of the gains (~65%) come from style over terminology.

Taken together, our work indicates that for the law, medical and Koran domains of the commonly-used multi-domain scenario, large LLMs need very little terminology help, and the improvements from demonstrations are more so from matching corpus style than from better conveying domain-specific semantics. Our work takes a first step in surfacing the domain-specific knowledge of smaller LLMs through generation, and we look forward to more informed approaches in future work. Meanwhile for the largest LLMs, we recommend as the most promising direction to construct a new MT adaptation scenario that challenges even their broad base of parametric knowledge, perhaps with reference to pretraining cut-off dates.

Limitations

While our work aims to generally study the problem domain adaptation for MT, we considered only a single dataset, multi-domain, with 3 domains and 1 language pair. This is following prior work, and also as there no other suitable datasets for our comparative setting. This is further discussed in Appendix A. We noted the limitations of this dataset, in not posing enough of a domain-adaptation challenge for current LLMs. We call on future work to design more up-to-date, comprehensive domain-adapted MT datasets.

We acknowledge that the multi-domain dataset is fairly well-worn, and there is possible data leakage into current LLMs, given the availability of the entire dataset online. This is a general concern with research using proprietary LLMs. However, the fact that demonstration retrieval does improve COMET scores for multi-domain indicates that, at the very least, the paired translations have not exactly memorized. Also, consider the Koran domain. While an LLM have undoubtedly seen Koran text during training, because there are multiple translations of the Koran into both English and in German, there is no exact 1-1 mapping with respect to the translations used in this dataset.

For demonstration retrieval, we used only the BM25 algorithm. Prior works have explored more informed retrieval approaches; however they were starting from much weaker zero-shot baseline, meaning that demo quality should matter less in our case. We reiterate that improving retrieval-based few-shot MT is not the goal of this work; rather, we aim to understand why it works well, and whether generating from parametric memory alone is viable. Our analysis, including our decomposition of demonstrations into style and terminology, can also be applied to demonstrations from any other similarity method.

Our use of a silver terminology built by LLM may lead to an under-estimation of the value of retrieved knowledge from bilingual terminology dictionaries. Likewise, our decomposition of demonstrations into terminology entries and style templates may be affected by the LLM's terminology-extraction errors. As mentioned in the main text, prior work indicates that these techniques (with older LLMs) should be roughly 95% accurate (Moslem et al., 2023).

7 Acknowledgments

We thank the members of Google Translate Research for their guidance throughout the course of this project. We thank Weiting Tan for providing us with the filtered version of the multi-domain test set, and the anonymous reviewers and area chair for their feedback.

References

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.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student*

Research Workshop, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *Preprint*, arXiv:2409.06790.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *Preprint*, arXiv:1612.06897.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

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.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3/4):251–282.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

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.

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*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor

Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Harold Somers. 1999. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Weiting Tan, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2024. Narrowing the gap between zero- and few-shot machine translation by matching styles. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 490–502, Mexico City, Mexico. Association for Computational Linguistics.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team. 2024b. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Shuo Wang, Peng Li, Zhixing Tan, Zhaopeng Tu, Maosong Sun, and Yang Liu. 2022. A template-based method for constrained neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3665–3679, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 993–1000, Manchester, UK. Coling 2008 Organizing Committee.

Masaru Yamada. 2011. The effect of translation memory databases on productivity. *Translation research projects*, 3:63–73.

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*, pages 41092–41110. PMLR.

A Other MT with Terminology Datasets

We did not use the datasets from WMT21 and WMT23 shared tasks on MT with terminologies. They do not include datastore for retrieving demonstrations, as well as each having its own concerns. For WMT21, we found that MT performance for zero-shot and using gold terms was equivalent (87.0 vs. 86.8 COMET22). This is due to contemporary LLM pretraining data containing a lot of COVID domain text, making it no longer a specialist domain. For WMT23, terminologies are internally defined – i.e., written directly with respect to each test and dev bitext. As we argued earlier, terminologies should be considered as external, pre-defined resources. We therefore recommend that both WMT21 and WMT23 datasets are outdated with current LLMs, and their use should be avoided.

Aycock and Bawden (2024) introduce a domain-adapted MT dataset, which they curate as a subset of existing MT resources from the OPUS project. This covers 7 domains and 11 languages. However, for all domains of their dataset, there is no large-scale data-store for demonstration retrieval; they only perform retrieval – proposing a topic-model guided exemplar selection method, which they show beats BM25 – over the very small development splits. Our work therefore considers only the multi-domain dataset, as it widely used for domain-adapted MT, and also satisfies our external resource requirements.

B Prompts Used

We reproduce the exact prompts used below, where `{<some_var>}` are variables which are filled per prompt, and `[<some_ex>]` are the static exemplars which are filled per-domain.

```
Instruction: Translate the following {src_full} text into {tgt_full} and output the result
in JSON format using "translation" as the key.
{source_language_name}: {source_text}
{target_language_name}:
```

Figure 4: Prompt for zero-shot MT.

```
You are tasked with translating {source_language_name} to {target_language_name}. You are provided
several example translations, and you should follow their example to translate the given
{source_language_name} sentence.
{demo_examples}
{source_language_name}: {source_text}
{target_language_name}:
```

Figure 5: Prompt for MT with *demonstrations* (also known as few-shot MT in prior work). This prompt is used for both demonstration retrieval and demonstration generation.

```
Your task is to translate a piece of text from {source_language_name} into {target_language_name}.
You are provided a list of terminology dictionaries. Each dictionary has a single source term (key
"de"), and multiple candidate translated terms (key "en") -- pick the most appropriate translated
term for the source sentence. Note that the terminologies have lowercased terms, but you should
consider proper casing when translating into {target_language_name}. Based on these terminologies,
output your best one translation.
{examples}
Terminology: {terminology}
{source_language_name}: {source_text}
{target_language_name}:
```

Figure 6: Prompt for MT with *terminologies*. This prompt is used for both terminology retrieval and terminology generation.

```
You are tasked with translating {source_language_name} to {target_language_name}. You are provided
several example translations, and you should follow their example to translate the given
{source_language_name} sentence. Note that the examples might contain special mask tokens <MASK> but
in your output, please do not use any such tokens.

[few_shot_examples]
{source_language_name}: {source_text}
{target_language_name}:
```

Figure 7: Prompt for MT with *style from demonstrations*. Recall that in this setting, we provide the retrieved demonstrations, but with the terminologies masked out – i.e., the style contribution is the inverse of the terminology contribution.

```
You are given a {source_language_name} source text, and asked to write exactly 3 text pairs. A text
pair consists of a {source_language_name} text, which is related to but different from the source
text, and its translation into {target_language_name}. You should do your best to ensure that your
{source_language_name} texts have similar style to the source text. Following the provided examples,
output each pair as a JSON dictionary, with keys "de" and "en". Each dictionary should be on a
separate line.
[demo_examples]
{source_language_name} source: {source_text}
Pair 1:
```

Figure 8: Prompt for synthetic *demonstration generation*. `[demo_examples]` are static exemplars for this task; see below.

```
German source: Die EDGE- und EDGE-II-Studien verglichen die gastrointestinale Verträglichkeit
von Etoricoxib mit der von Diclofenac.
Pair 1: {"de": "Die kardiorenalen Ergebnisse der EDGE- und EDGE-II-Studien entsprachen den
für die MEDAL- Studie beschriebenen.", "en": "The cardiorenal results for EDGE and EDGE II
were consistent with those described for the MEDAL Study."}
Pair 2: {"de": "Eine langsame Dosissteigerung kann die gastrointestinale Verträglichkeit
ebenfalls verbessern.", "en": "A slow increase in the dose may also improve gastrointestinal
tolerability."}
Pair 3: {"de": "Die Häufigkeit von unerwünschten Ereignissen in EDGE und EDGE II sowie die
Häufigkeit von als schwerwiegend erachteten oder zum Studienabbruch führenden unerwünschten
Ereignissen in der MEDAL-Studie war unter Etoricoxib höher als unter Diclofenac.", "en": "The
incidence of adverse experiences in EDGE and EDGE II and of adverse experiences considered
serious or resulting in discontinuation in the MEDAL study was higher with etoricoxib than
diclofenac."}
---
German source: 3 ml Lösung in einer Patrone aus Glas (Glasart 1), mit einem Kolben (Brombutylgummi)
und einem Stopfen (Brombutylgummi/Polyisopren) in einem Umkarton.
Pair 1: {"de": "3 ml Lösung in einer Patrone aus Glas (Glasart 1), mit einem Kolben (Brombutylgummi)
und einem Stopfen (Brombutylgummi/Polyisopren) in einem Fertigpen (Mehrdosen-Einwegspritze aus
Polypropylen).", "en": "3 ml solution in a cartridge (type 1 glass) with a plunger (bromobutyl) and
a stopper (bromobutyl/ polyisoprene) contained in a pre-filled pen (multidose disposable pen)
(polypropylene)."}
Pair 2: {"de": "3 ml Suspension in einer Patrone (farbloses Glas, Typ 1) mit einem Kolben
(Brombutylgummi, Typ 1) und einer Bördelkappe (Aluminium) mit einem Stopfen (Brombutyl- oder
Polyisopren- Brombutylgummi, Typ 1).", "en": "3 ml suspension in a cartridge (type 1 colourless
glass) with a plunger (bromobutyl rubber (type 1)) and a flanged cap (aluminium) with a stopper
(bromobutyl or laminate of polyisoprene and bromobutyl rubber (type 1))."}
Pair 3: {"de": "5 ml Lösung in einer Durchstechflasche (farbloses Glas, Typ 1) mit einer
Bördelkappe (Aluminium), einem Stopfen (Chlorbutylgummi, Typ 1) und einem Abreißdeckel
(Polypropylen).", "en": "5 ml solution in a vial (type 1 colourless glass) with a flanged cap
(aluminium), a stopper (chlorobutyl rubber (type 1)) and a tear-off cap (polypropylene)."}
---
```

Figure 9: Static 2-shot exemplars used for the synthetic *demonstration generation* prompt (Figure 8). Each exemplar has 3 output sentences. Here we show the exemplars for the medical domain.

```
You are given a {source_language_name} source text, and asked to extract a bilingual terminology
that translates key terms from the source text into {target_language_name}. Each entry in the
terminology should have a {source_language_name} term and a list of possible {target_language_name}
translations. Following the provided examples, output each pair as a JSON dictionary, with keys "de"
and "en". Each dictionary should be on a separate line.
[term_examples]
{source_language_name} source: {source_text}
Term 1:
```

Figure 10: Prompt for synthetic *terminology generation*. `[term_examples]` are static exemplars for this task; see below.

```
German source: (6) Die Kommission unterrichtete den Antragsteller, andere Gemeinschaftshersteller,
die ausführenden Hersteller in der VR China und in den USA, bekanntermaßen betroffene Einführer
und Verwender sowie die Vertreter der Regierungen der VR China und der USA offiziell über die
Einleitung des Verfahrens.
Term 1: {"de": "einleitung des verfahrens", "en": ["initiation of the proceeding",
"opening of the proceedings"]}
Term 2: {"de": "ausführenden hersteller", "en": ["exporting producers"]}
Term 3: {"de": "gemeinschaftshersteller", "en": ["community producers"]}
Term 4: {"de": "antragsteller", "en": ["complainant"]}
Term 5: {"de": "kommission", "en": ["commission"]}
Term 6: {"de": "verfahrens", "en": ["investigation", "procedure"]}
Term 7: {"de": "einführer", "en": ["importers"]}
Term 8: {"de": "verwender", "en": ["users"]}
Term 9: {"de": "vertreter", "en": ["representatives"]}
Term 10: {"de": "vr china", "en": ["prc"]}
---
German source: ENTSCHEIDUNG DER KOMMISSION vom 25. Februar 1998 zum Fragebogen für die Berichte der
Mitgliedstaaten über die Umsetzung der Richtlinie 94/67/EG des Rates über die Verbrennung
gefährlicher Abfälle (Umsetzung der Richtlinie 91/692/EWG des Rates) (Text von Bedeutung
für den EWR) (98/184/EG)
Term 1: {"de": "verbrennung gefährlicher abfälle", "en": ["incineration of hazardous waste"]}
Term 2: {"de": "fragebogen", "en": ["questionnaire"]}
Term 3: {"de": "richtlinie", "en": ["directive", "guideline"]}
Term 4: {"de": "ewr", "en": ["eea relevance"]}
---
```

Figure 11: The static 2-shot exemplars used for the synthetic *terminology generation* prompt (Figure 10). Here we show the exemplars for the law domain.

```
Identify and annotate all terminology entities (consider only consecutive words) from the source
sentences and match them with the counterpart in the target sentences. Your response should follow
the format of the provided examples, so that each numbered source and target pair corresponds to
exactly one terminology line in your response.
[source_examples]
{source_texts}
---
[target_examples]
{target_texts}
---
[term_examples]
```

Figure 12: Prompt for *terminology extraction* from source-target text pairs. For each prompt, we batch together 5 text pairs to extract from at a time. `[source_examples]`, `[target_examples]`, `[term_examples]` are static exemplars for this task; see below.

```
source 1: Sag: "Wer hat denn die Schrift hinabgesandt, mit der Musa als Licht und als Rechtleitung
für die Menschen kam?
source 2: Sollte Seine Peinigung über euch nachts oder am Tage hereinbrechen, was wollen denn die
schwer Verfehlenden davon beschleunigen?"
source 3: Unser Herr! Du bist wahrlich Gütig, Barmherzig."
source 4: Und diejenigen, die an Allah und Seine Gesandten glauben, sind die Wahrhaftigen und die
Bezeugenden vor ihrem Herrn; sie werden ihren Lohn und ihr Licht empfangen.
source 5: "Wer sich im Irrtum befindet, dem soll Der Allgnade Erweisende noch mehr davon gewähren!"
Wenn sie dann sehen, was ihnen angedroht wurde: entweder die Peinigung oder die Stunde, dann werden
sie wissen, wer über die schlimmere Stellung und die schwächere Streitmacht verfügt.
```

```
target 1: Say: "Who sent down the Book that Moses brought as a light and a guidance to men?
target 2: If His chastisement comes upon you by night or day, what part of it will the sinners seek
to hasten?
target 3: Our Lord, surely Thou art the All-gentle, the All-compassionate."
target 4: Those who believe in God and His apostles are true of word and deed; and by their Lord are
considered testifiers of the truth. They have their guerdon and their light.
target 5: "Ar-Rahman extends the life of those who are astray until they come to realise what had
been promised them was either (physical) affliction or (the terror) of Resurrection. Then will they
know who is worse in position, and who is weak in supporters.
```

```
terminology 1: [{{"en": "Book", "de": "Schrift"}}, {{"en": "guidance", "de": "Rechtleitung"}},
{{"en": "Moses", "de": "Musa"}}]
terminology 2: [{{"en": "chastisement", "de": "Peinigung"}}, {{"en": "sinners", "de": "schwer
Verfehlenden"}}]
terminology 3: [{{"en": "Our Lord", "de": "Unser Herr"}}, {{"en": "All-gentle", "de": "Gütig"}},
{{"en": "All-compassionate", "de": "Barmherzig"}}]
terminology 4: [{{"en": "His apostles", "de": "Seine Gesandten"}} {{"en": "true of word and deed",
"de": "die Wahrhaftigen und die Bezeugenden"}}, {{"en": "by their Lord", "de": "vor ihrem Herrn"}},
{{"en": "their guerdon", "de": "ihren Lohn"}}, {{"en": "their light", "de": "ihr Licht"}}]
terminology 5: [{{"en": "Ar-Rahman", "de": "Der Allgnade Erweisende"}}, {{"en": "extends the life",
"de": "noch mehr davon gewähren"}}, {{"en": "those who are astray", "de": "Wer sich im Irrtum
befindet"}}, {{"en": "come to realise", "de": "sehen"}}, {{"en": "promised", "de": "angedroht"}},
{{"en": "(physical) affliction", "de": "Peinigung"}}, {{"en": "(the terror) of Resurrection", "de":
"Stunde"}}, {{"en": "worse in position", "de": "über die schlimmere Stellung"}}, {{"en": "weak in
supporters", "de": "die schwächere Streitmacht"}}]
```

Figure 13: Static 5-shot exemplars used for the synthetic *terminology extraction* prompt (Figure 12). We found that this format, where each of the 3 blocks (source, target, terms) are consecutive to each other, gave the most parseable output. Note that the 5 exemplars is the same size as the batches of 5 to extract terminologies from. Here we show the exemplars for the koran domain.