In-context Learning as Maintaining Coherency: A Study of On-the-fly Machine Translation Using Large Language Models

Suzanna Sia
Johns Hopkins University
ssia1@jhu.edu

Kevin Duh
Johns Hopkins University
kevinduh@cs.jhu.edu

Abstract

The phenomena of in-context learning has typically been thought of as "learning from examples". In this work which focuses on Machine Translation, we present a perspective of in-context learning as the desired generation task maintaining coherency with its context, i.e., the prompt examples. We first investigate randomly sampled prompts across 4 domains, and find that translation performance improves when shown in-domain prompts. Next, we investigate coherency for the in-domain setting, which uses prompt examples from a moving window. We study this with respect to other factors that have previously been identified in the literature such as length, surface similarity and sentence embedding similarity. Our results across 3 models (GPTNeo2.7B, Bloom3B, XGLM2.9B), and three translation directions (en→{pt, de, fr}) suggest that the long-term coherency of the prompts and the test sentence is a good indicator of downstream translation performance. In doing so, we demonstrate the efficacy of In-context Machine Translation for on-the-fly adaptation.

1 Introduction

In-context Machine Translation is a relatively new paradigm that uses large autoregressive Language Models to carry out the task of Machine Translation (MT) by being shown translation pairs in the prefix. From a practitioner's viewpoint, In-context learning presents itself as an attractive approach for rapidly adapting a Translation model on-the-fly. Previous strategies for adapting a pre-trained MT model still require additional engineering or training of the model, e.g fine-tuning with in-domain data using adaptor layers (Philip et al., 2020). Instead, simply changing the inputs to the model might be an effective way to adapt on-the-fly without any model modification.

The in-context learning paradigm describes a phenomena where large autoregressive language models perform a task when shown examples (known as prompts) in the prefix (Brown et al., 2020; Bommasani et al., 2021). Previous work approaches the role of the prompt context as allowing the model to "learn by examples". This intuitive approach to formulating the task of prompt selection has led to the suggestion of selecting examples that are similar to the source sentence being translated. Semantic similarity based on sentence embeddings (Liu et al., 2021) and BM25 have been proposed to select examples to present as "demonstrations" (Rubin et al., 2021). This approach was further expanded by Agrawal et al. (2022) who show that BM25 and a heuristic version optimizing for word coverage, is effective for selecting examples.

We focus on Machine Translation as a complex conditional generation task and offer an alternate perspective: **the in-context paradigm depends on maintaining *coherency*.** Coherence is an aspect of natural language that reflects the overall semantic and syntatic consistency in a body of text (Flowerdew and Mahlberg, 2009). We investigate this by first exploring the model's behavior when showing matching and mismatching domains in the context and the test sentence. Next we consider a stricter notion of coherency using a moving window of previous gold translations directly preceding the test source sentence to be next translated. Our experiments compare the coherence factor with similarity based factors for prompt selection, additionally controlling for length (Xie et al., 2021) which is typically overlooked but is important to consider for performance and available labeling (translation) budget. The contributions of this work are

- We identify coherency of prompt examples with respect to test sentence as a critical factor for translation performance. Experiments across 3 models (GPTNeo2.7B, Bloom3B, XGLM2.9B) and 4 domains (Medical, Social Media, Wikipedia, and TED Talks) suggest that mod-

Translate English to French.		
English: A discomfort which lasts ..	French:	Un malaise qui dure
English: HTML is a language for formatting	French:	HTML est un langage de formatage
...		...
English: After you become comfortable with formatting ..	French:	

Table 1: A single continuous input sequence presented to the model for decoding a single test source sentence "After you become comfortable with formatting..". Given the entire sequence as input, the model proceeds to generate the target sequence.

els perform better when prompts are randomly drawn from the same domain.

- Within the TED talks domain, we investigate local coherence using document-level translation experiments, by adopting a moving window directly preceding the test source sentence to be translated. Overall, our results across the 3 models and three translation directions (en→{pt, de, fr}) suggest that the coherency of the prompts with regard to the test sentence is a good indicator of translation performance.

2 Preliminaries

2.1 In-context Machine Translation

In an in-context learning setup, several formatting decisions need to be made on how to present the prompt examples to the model. We adopt the following commonly used prompt format where the instructions are straightforwardly provided as in the following (Table 1).[1] In this work, we consider both sentence level translation (Section 5.1) and an on-the-fly document-level setting (Section 5.3).

2.2 Coherence in Natural Language Text

The computational linguistics literature holds many competing definitions of coherence in text (Wang and Guo, 2014). We consider two aspects of coherence, first from a more global level where we investigate domain effects, and also from a local sentence level, where we consider a coherent context as a moving window of previous (gold) translations which directly precede a test sentence. A similar working definition of coherence has been used in discrimination tasks that require a model to identifying the right order of (shuffled) sentences (Elsner et al., 2007; Barzilay and Lapata, 2008; Laban et al., 2021).

[1]We also experiment with a different separator "=" used in (Lin et al., 2021) (instead of "English" and "French"), but find that this does not perform significantly better.

3 Factors which affect In-context MT

We outline several factors studied in this paper related to example selection for In-context MT. While we emphasise the notion of *Coherence* (Section 2.2), by studying the domain factor (Section 3.4) and local coherence (Section 3.5), our experiments seek to compare this against other factors that have been highlighted in previous literature. Namely, length (Section 3.1), surface similarity (Section 3.2) and semantic similarity (Section 3.3). To demonstrate, in Table 1, the first sentence is semantically similar and the second sentence has surface similarity with the test sentence.

3.1 Length (Translation Budget)

One previously overlooked factor for in-context MT is the length (number of words) in the prompt examples. The perspective of In-context Learning as implicit Bayesian Inference argues that longer examples provide more evidence to the model on the desired task pattern (Xie et al., 2021). Longer examples are also more likely to contain non-trivial translation exemplars, although it is not clear whether this affects downstream performance. We find example length to be correlated with the domain (Figure 2), and it may thus be a confounding factor for in-context MT.

Controlling for Length We adopt the notion of a "Translation Budget" which is the total word count of all the prompt examples provided (excluding the test sentence). Examples can be selected as long as they satisfy the budget constraint. A generalized algorithm is provided in Section 4.3. From a resource perspective, this reflects the work of the human annotator in providing example translations.

3.2 Surface Similarity

3.2.1 BM25

BM25 (Robertson et al., 2009) is a bag-of-words unsupervised retrieval function that ranks a set of documents based on the query terms appearing in the documents. Agrawal et al. (2022) report that using BM25 to retrieve similar prompt examples

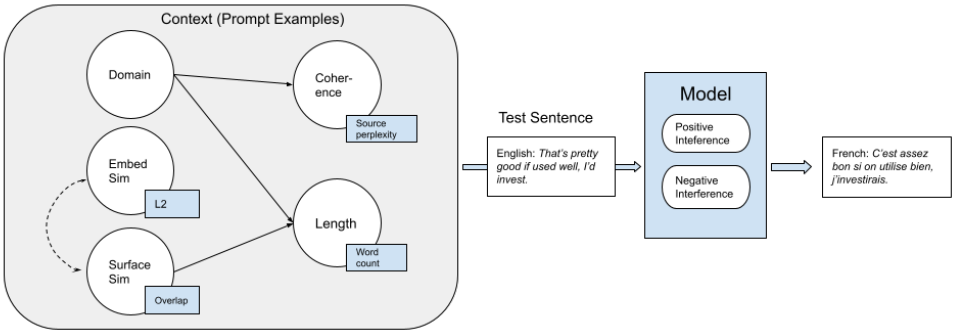

Figure 1: Factors identified and studied in this paper. Selecting from matching Domain increases coherence (Appendix C) and each domain has different length distributions (Section 5.2). Surface similarity and embedding similarity are associated (Table 4). Surface similarity selection also results in longer sentences (Section 5.4) Rectangle boxes next to the node are measures of these factors. We describe and quantify positive and negative interference of the model for translation performance in Section 6.3.

outperforms random selection. They also advocate for a variant of BM25 with increased coverage of test sentence source words although with marginal gains (<1 BLEU point) increase. Following Agrawal et al. (2022), we order the examples according to their similarity to the source, with the most similar examples on the left in all our experiments.

3.2.2 Maximising Surface Similarity Coverage

To maximise word overlap across all prompts and the source sentence, we adopt Submodular optimisation by Maximal Marginal Relevance (Carbonell and Goldstein, 1998; Lin and Bilmes, 2010). Formally we are given a finite size set of objects U (the size of the prompt bank). A valuation function $f : 2^U \rightarrow \mathcal{R}_+$ returns a non-negative real value for any subset $X \subset U$. The function f is said to be submodular if it satisfies the property of "diminishing returns", namely, for all $X \subset Z$ and $Z \notin U$, we have $f(X \cup u) - f(X) \geq f(Z \cup u) - f(Z)$. The algorithm optimises for sentences with maximal word overlap weighted by the BM25 score.

3.3 Semantic Similarity (Nearest Neighbors)

The semantic similarity of prompts based on their sentence embeddings has also been advocated for selecting good in-context examples. Liu et al. (2021) apply a pre-trained Roberta-large sentence encoder to the test sentence, and query for its nearest neighbors to use as in-context demonstrations. In our experiments we apply a similar strategy using MPNet base (Song et al., 2020) which achieved highest scores on HuggingFace sentence embedding and semantic search benchmarks.[2] We do not consider training a prompt retriever (Rubin et al., 2021) or fine-tuning the sentence encoder (Liu et al., 2021) in this study, as these are no longer "light-weight" retrieval methods that are comparable with the other unsupervised strategies.

3.4 Domain Coherence

GPT is able to do style transfer just from instructions or from being shown surface prompt examples (Reif et al., 2022). Simply providing demonstrations from the same domain may induce the large language model (LLM) to generate a similar style which is coherent with the target text. Another possibility is that particular lexical translation exemplars which match the source sentence may be present. However, due to the very high dimensionality of the raw vocabulary, this is less likely if translation examples are randomly sampled.

Domain may also present spurious correlations which are confounded by the training data of LLMs. For instance, there may be certain domains which are better at eliciting Translation behavior from the model, regardless of what the test domain is.

3.5 Local Coherence (Moving Window)

We hypothesise that the local coherence (Section 2.2) of the context to the test sentence to be translated may be an important factor for performance. To test this, we adopt a moving context window of the previously translated gold sentence pairs as the prompt examples. To our knowledge, Section 3.4 and Section 3.5 are previously unexplored for In-context Machine Translation.

[2] https://www.sbert.net/docs/pretrained_models.html

4 Experiments

4.1 Data

Domain Coherence We organise our experiments investigating four en→fr domains, WMT19 Biomedical (MED) (Bawden et al., 2019), a social media dataset, MTNT (Michel and Neubig, 2018), multilingual TED Talks, and Wikipedia-based FLORES (Goyal et al., 2021). Except for MED, all other datasets have a wide range of topics in the train (prompt bank) and test set which are shuffled in random sampling, and thus the domain experiments are more focused on the writing style of the text.

We use standard train-test splits, with the trainset being used as the prompt bank. Scores are reported using SacreBLEU (Post, 2018).[3]

Local Coherence (document level) We use the Multitarget TED Talks dataset from Duh (2018). The original dataset has 30 documents in the test set, where each document corresponds to a 10-20 minute TED talk. To increase the size of the test set, we partition the "original" trainset into a train (prompt bank) and test split, where talks with a minimum of 100 lines were used as the test and talks with less than 100 lines were used as the "out-of-document" prompt bank. We used 120 test documents that had a minimum of 100 lines, and we evaluated each up to 120 lines, where each TED talk is a document. The document level BLEU scores are reported for three language directions en→{fr, pt, de}. We do not use a dev set as there is no training or any tuning of any hyperparameters. Since this is a non-standardised data split, we provide the numbers in the following table.

	Talks (Docs)	Lines per doc	Total Lines
"Outside-doc" Prompt Bank	450	<100	26000+
"Within-doc" Prompt Bank	1	100-120	120
Test	120	100-120	12000+

4.2 Models

We use three models, GPTNeo2.7B (Black et al., 2021), XGLM2.9B (Lin et al., 2021), and Bloom3B (Scao et al., 2022) which are open access LLMs available on HuggingFace (Wolf et al., 2020). The later two have been advertised as "Multilingual Language Models". GPTNeo2.7B is a GPT3 replicate pretrained on The Pile (Gao et al., 2020), while XGLM adopts a similar architecture trained on a multilingual corpus (CC100-XL). Bloom3B has been trained on the ROOTS Corpus (Laurençon et al., 2022), a collection of huggingface datasets of 1.6 TB of text. To our knowledge, there has not been any reports of sentence level parallel corpora in the training datasets of these models.

[3] nrefs:1 | case:lower | eff:no | tok:13a | smooth:exp | version:2.0.0

4.3 Algorithm for Greedy selection with Length Constraint

In our experiments, we investigate BM25 (Section 3.2.1), BM25 with submodular optimisation (BM25-s; Section 3.2.2), and semantic similarity (nn; Section 3.3). To control for length effects, we employ an algorithm for selection with length constraints (algorithm 1) which closely follows greedy submodular algorithms (Krause and Guestrin, 2008). Retrieval methods adopts a utility function: f, which is used to retrieve highest scoring sentences. For BM25 and BM25-s, $f$is BM25, while u_i is selected by $f(\{u\})$, and $f(\{u\}|X_i)$ respectively. While for nn, f is the L2 embedding similarity between prompt sentence and test query.

Algorithm 1: Generalised greedy (submodular) algorithm with length budget

Input: (Submodular) function $f : 2^U \rightarrow R_+$, cost function m, budget b, finite prompt bank U
Output: X_k where k is the number of iterations/prompts.
Set $X_0 \leftarrow$; $i \leftarrow 0$;
while $m(X_i) < b$ **do**
  $u_i = \text{argmax}_{u \in U \setminus X_i} f(\{u\} \mid X_i)$
  $X_{i+1} \leftarrow X_i \cup u_i$;
  $i \leftarrow i + 1$

5 Analysis of Factors

5.1 Domain Coherence [Table 2]

Does coherence of domain allow models to adapt on the fly? If models are adapting to the domain shown in the context, sampling and testing within the same domain should result in the highest translation performance, as compared to being shown examples out of domain. For example, if we are

	GPTNeo2.7B				Bloom3B				XGLM2.9B			
Prompt / Test	FLORES	MED	MTNT	TED	FLORES	MED	MTNT	TED	FLORES	MED	MTNT	TED
FLORES	**24.6**	**19.7**	**23.1**	**24.6**	**36.7**	28.5	28.5	31.1	**29.3**	20.9	24.7	**25.7**
MED	23.0	19.2	21.1	23.2	34.5	**28.7**	26.2	29.5	27.5	**21.4**	22.9	24.4
MTNT	23.7	18.6	22.4	23.7	35.5	27.7	**29.1**	30.6	27.9	21.2	**25.0**	25.4
TED	23.2	18.6	22.1	23.6	36.1	27.9	**29.1**	**31.2**	27.8	21.1	24.2	24.8

Table 2: Crosstable of BLEU scores from sampling and testing in different domains. We present the average BLEU scores across 5 randomly sampled prompt sets. The size of the prompt sets (number of translation pair examples) is 5. We bold the largest value column-wise.

testing on the TED domain, is it important that the prompt be also drawn from TED or is it sufficient to have sentence pairs from any domain illustrating the translation task? To account for prompt selection and ordering effects, all inference runs were repeated with 5 randomly sampled prompt sets from the training data. We focus on en → fr which is common across datasets.

Results and Discussion

- Models are able to perform some form of domain adaptation on-the-fly. There appears to be evidence of domain adaptation in Bloom3B and XGLM, as sampling and testing within the same domain (e.g., sample from MED test with MED) mostly results in the highest performance column-wise. We also observe that matching domains result in lower conditional source sentence perplexity (Appendix C).

- For GPTNeo, sampling from FLORES results in the best translation performance across all test sentences even with domain mismatch. This suggests that translation performance in GPTNeo is best induced using FLORES and is less adaptive to the domain. Note that the second best column wise result for GPTNeo tends to occur when there is matching prompt and test domain.

5.2 Domain controlling for Length

How does length of prompts affect translation across different domains? In Figure 2, we randomly sample 1000 sentences from each domain's training set. Randomly sampled sentences from different domains show distinct length effects. We study the impact of these length effects by selecting either a 5-10 word or 15-20 word long sentences for translation examples, and compare the differences in scores for the non-filtered scenario (Table 3).

Results and Discussion

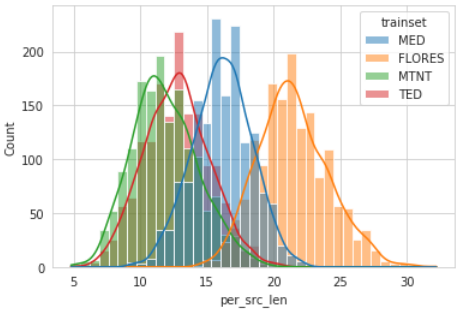

Figure 2: Histograms of sentence lengths (word counts) randomly sampled from different domains, which has implications for the total prompt length when sampling from these domains. FLORES sentences tend to be nearly twice as long as MTNT and TED sentences.

Prompt / Test	FLORES	MED	MTNT	TED
FLORES	-	-	-	-
MED	⇓22.4 (0.3)	⇓18.5 (0.3)	↓20.8 (0.8)	⇓22.5 (0.8)
MTNT	⇓23.2 (0.4)	↓18.3 (0.5)	⇓21.9 (1.2)	↓23.5 (0.5)
TED	⇓21.7 (1.4)	⇓17.6 (0.6)	⇓20.1 (1.8)	⇓22.3 (1.5)

5-10 words long sentences; GPTNeo 2.7B

Prompt / Test	FLORES	MED	MTNT	TED
FLORES	24.2 (0.2)↓	19.6 (0.3)	22.7 (0.8)↓	24.3 (0.5)↓
MED	22.9 (0.6)	19.3 (0.1)	21.1 (0.9)	22.8 (0.7) ↓
MTNT	24.0 (0.4) ↑	18.9 (0.6)↑	22.5 (0.0)	24.3 (0.3)⇑
TED	23.8 (0.4)⇑	19.0 (0.4)↑	22.9 (0.2)⇑	23.8 (0.4)

15-20 words long sentences; GPTNeo 2.7B

Table 3: Selecting for short source sentences (5-10 words) vs longer source sentences (15-20 words) as translation examples. ↓and ↑refers to differences > 0.3, and ⇓and ⇑refers to differences > 0.5 when compared to the no-length filter scenario in Table 2.

- When source prompt sentences are 5-10 words, all BLEU scores decrease. For 15-20 words sentences which is "long" for MTNT and TED, but "short" for FLORES, the BLEU score of the former increases while the latter decreases. BLEU scores are similar for MED as 15-20 words is close to the mean of MED length distribution.

- We inspect the length of generation under different prompt lengths, and find that average differences in generation length are marginal (only 1-2 words difference) indicating that poorer per-

formance is not simply due to a difference in generation lengths.

5.3 Local Coherence [Table 4]

How important is a coherent context (as compared to other prompt selection methods?) Section 5.1 showed that models are able to adapt when shown prompts from a matching domain. We hypothesise that coherence of the prompts with respect to the test source sentence (Section 2.2) is an important factor for performance.

We use the TED talks dataset (data preparation described in Section 4.1), and consider a moving window of previous gold translations (`window`) as a coherent context for the model.[4] We compare this against the baselines of (`BM25`; Section 3.2.1), (`BM25-s`; Section 3.2.2), and Nearest Neighbor retrieval of sentence embeddings (`nn`; Section 3.3) from a large prompt bank outside the document. We use a prompt set of 5 examples for all experiments, and randomly sample from outside of the document if the available window is smaller than 5. Document level BLEU scores are averaged across 120 documents and reported in Table 4.

Quantifying Similarity We report the ROUGE1-precision (`coverage`; Lin (2004)) and the L2 Euclidean distance (`L2`) of the source sentences in the prompt set, with the test source sentence to be translated. If translation performance is due to word overlap or embedding similarity, then we expect that having a higher `coverage` or lower `L2` would have better performance than `window`. Note that all similarity based retrieval methods depend only on the source sentences, and is model and target language independent. i.e., the single `coverage` and `L2` value applies for all results columns in Table 4.

Results and Discussion

- The moving window (`window`) outperforms all other baselines across the 3 models and 3 language directions, with the exception of Bloom3B on `en`→`de` direction. The gains are from 0.5 to 2.6 BLEU points from the next best performing retrieval method. Importantly, `coverage` and `L2` shows that the performance is not due to similarity or word overlap.

- Interestingly, randomly sampling sentences from within the document (talk) performs well compared to other similarity based retrieval methods from outside of the document. This further highlights that coherence is a critical factor for In-context Machine Translation.

- Similarity based retrieval mostly does better than randomly sampled prompt sets, which is consistent with existing literature which did not consider the factor of coherence. A notable exception is XGLM `en`→`fr` results, where similarity based methods are doing poorly compared to that reported by (Agrawal et al., 2022). We find that the similarity based retrieval methods does better for XGLM when the number of prompts is increased from 5 to 15. The same trend is observed at 15 prompts, `window` continues to outperform the other methods (results in Appendix B).

Crucially, this set of experiments show that *similarity based methods are not as critical for translation as compared to coherency*, a new factor that we identify in this work.

5.4 Similarity based Retrieval within the Document

How well do similarity based retrieval methods perform for previous on-the-fly translations? In Section 5.3, we established that using a moving window (local coherence) outperforms retrieval from outside the document with similarity-based retrieval methods. Here we apply `bm25`, `bm25-s`, `nn` for retrieval *within* the document. We consider the more realistic "on-the-fly" or computer-aided translation scenario, where the human translator works with MT systems, and translation examples in the document can only be selected prior to the test sentence (Alabau et al., 2014).

Controlling for Length When doing retrieval based methods within the document for an "on-the-fly" setting, length factors in and longer sentences are retrieved on average. We thus investigate budgeting for the length constraint to be same as the moving window (`window`). For every test sentence, we compute the budget used by it's own moving window, and apply it as a length constraint to for the other retrieval based methods as described in Section 4.3. Results are presented in Figure 3.

Results and Discussion

[4] Preliminary experiments using model generated instead of gold translations performed worse than random.

		GPTNeo2.7B(BLEU)			Bloom3B(BLEU)			XGLM2.9B(BLEU)			L2	coverage
	In/outdoc	en→fr	en-pt	en-de	en-fr	en-pt	en-de	en-fr	en-pt	en-de	-	-
random	out	26.3	27.1	16.6	35.2	35.5	7.9	27.1	26.7	18.9	1.35	0.31
nn	out	26.8	26.9	16.9	35.1	35.1	8.2	25.6	26.6	18.3	0.98	0.49
bm25	out	27.1	27.4	17.3	35.1	35.3	**9.4**	25.3	27.0	18.4	1.21	0.75
bm25-s	out	27.2	27.5	17.4	34.8	34.9	9.1	25.6	27.4	18.7	1.25	0.80
random	within	27.4	27.3	17.3	35.9	35.8	7.8	26.6	28.8	19.6	1.28	0.34
window	within	**28.1**	**28.3**	**17.9**	**36.9**	**37.0**	8.8	**28.6**	**31.6**	**21.2**	1.22	0.40

Table 4: BLEU score comparison of similarity-based retrieval methods from out of document, and moving window (window) from within the document. Coverage (Rouge1-precision) refers to the word overlap between prompt source sentences and test source sentence. L2 refers to the average L2 Euclidean distance between source prompt sentence embeddings and the test sentence embedding.

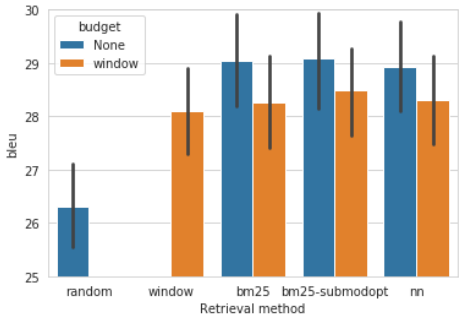

Figure 3: Comparison of Retrieval methods controling for budget: No budget or same budget as moving window. Model is GPTNeo2.7B on en->fr. random is sampled within the document.

- We observe similar performance for all retrieval methods, with bm25-s doing slightly better than bm25 and nearest neighbors (nn).

- Without any budget restriction, performance of retrieval methods outperforms window. However when restricted to the same budget as window, we find that the performance is within 0.1-0.5 BLEU score difference. Furthermore, the coverage is only 0.01-0.03 less if not using similarity based retrieval, indicating that most of the differences in contributions could be coming from the length effect and not because of similarity.

6 Further Analysis and Discussion

In this section, we focus on GPTNeo2.7B and in the en→fr direction.

6.1 Perplexity and Coverage

One natural question that arises is the relationship between Coverage, Coherence, and translation performance. Although there is no widely accepted measure of *general coherence*, we can formulate this with respect to the particular model being studied. We consider the model's conditional perplexity

retrieval	bleu	L2	Coverage	ppl_s
static	26.6	1.22	0.41	16.8
random	27.4	1.28	0.31	14.9
window	28.1	1.22	0.40	11.1
shuffle	28.3	1.22	0.40	12.0

Table 5: Ordering effects within document. All retrieval methods are within documnent.

of the test sentence given the context. Perplexity is a widely used measure of suprisal in text and has also been used as a measure in topic coherence (Newman et al., 2010). Concurrent work by Gonen et al. (2022) argue that total perplexity of the input sequence is related to In-context performance.

In Figure 4, we produce scatterplots of Sentence BLEU scores, source perplexity and Coverage (word overlap). We observe that there is a negative relationship between source perplexity and Sentence BLEU (-0.22 Pearson's r), but very noisy relationship between Sentence BLEU and word overlap, and word overlap and source perplexity.

6.2 Studying Local Coherence [Table 5]

We compare the window with other baselines which may give some indication of what is important in the document in terms of local coherence.

- Shuffle simulates whether the model is affected by the the local coherence by shuffling sentences within window.

- Static refers to the first k (window size) translation sentences of the document which is then held fix throughout when translating the rest of the document.

Interestingly, shuffling the set of prompts within the moving window which breaks the natural ordering of the document "coherence" does not deteriorate in-context translation performance. This

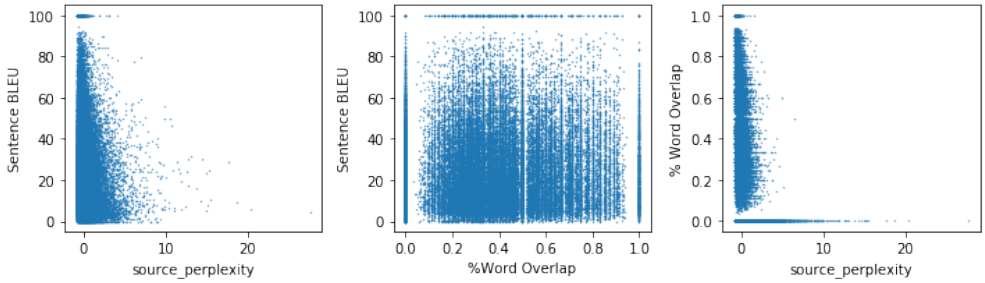

Figure 4: Scatterplots of Sentence BLEU Scores, with Source Perplexity and Word Overlap

	random	bm25	bm25-s	nn	window
Positive	0.56	**0.62**	0.61	0.6	**0.62**
Negative	0.32	0.31	0.31	0.32	**0.29**
No Change	0.12	0.07	0.08	0.08	0.09

Table 6: Positive, Negative and No change (proportions) in BLEU scores across different prompt selection methods. For positive row, higher is better. For negative row, lower is better.

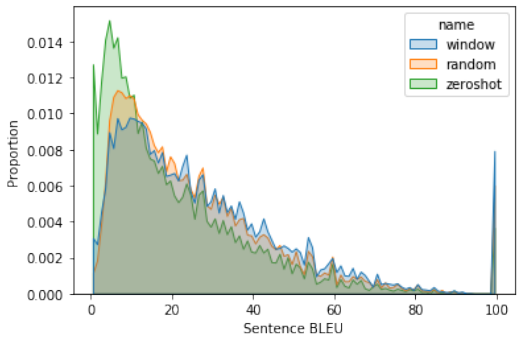

Figure 5: Histograms of sentence BLEU scores for `zeroshot` (no prompts), `random`, and `window`.

finding is consistent across several models and languages Appendix D. The ordering of the document does affect source perplexity, with perplexity increasing from 11.1 → to 12.0, however this does not negatively affect translation performance. This suggests that the relationship between coherence and translation is indirect or non-linear, and the way models use context might be counter-intuitive; a view increasingly advocated by recent research (Webson and Pavlick, 2021; Min et al., 2022). Overall this suggests we may benefit from methods which perform selection from within the document which we leave to future work.

6.3 Do we need Translation examples at all?

Given the rise of instruction-following GPT (Ouyang et al., 2022) a reasonable question is whether prompt example selection will still be relevant in future models. For a large language model, merely providing the instruction "Translate English to French" without any prompt examples (zero-shot) can still elicit a translation. In spite of zero-shot success, a common finding (for MT as well as other NLP tasks) is that providing more prompt examples typically results in better performance albeit with diminishing effect. Since examples are not strictly *necessary* for translation but can *enhance* the model's downstream translation ability, what is the role of prompt examples?

Positive vs Negative Task Interference One curiosity that we observe across all of our experiments, is that prompt sets *do better on-average rather than across all examples*, relative to the Zero-shot, instructions only setting. This suggests a notion of interference; examples may guide generation towards a poorer translation (negative interference) or better translation (positive interference). A closely related concept is task location (Reynolds and McDonell, 2021).

Table 6 quantifies this across different methods corresponding to the results in Table 4 for GPT-Neo2.7B en→ fr direction. `window` has both the highest positive interference and lowest negative interference. From Figure 5, the major role of prompting methods compared to the zero-shot scenario is to have greater positive interference chiefly over sentence BLEU of 20-60 and for some extreme cases of 100 BLEU, although a large proportion of sentences still lie in the low-scoring region.

7 Conclusion

In-context Learning has typically been thought of as learning from examples. In this work, we introduce a different perspective of coherency of the context with the test sentence. We first showed that models are mostly able to adapt to different writing styles when the prompt bank and test set are matching/consistent in domain. Experiments across 3 models and 3 languages show that a moving win-

dow is up to 2.6 BLEU points better than previously reported similarity based retrieval methods from outside the document. From this perspective, the problem of prompt selection for in-context MT is one of maintaining a coherency for text generation. Preliminary analysis on local coherence effects, and the presence of negative interference compared to the zero-shot setting, suggests avenues for future work on investigating more careful mechanisms for controlling in-context Machine Translation.

8 Limitations

This section details several limitations and ethical concerns associated with this work.

- While we have identified coherency of domain and document as a factor for in-context MT, we expect there should be other factors that could be more predictive of downstream performance, such as activation of attention patterns from source to target sentence during generation.
- We studied GPTNeo, Bloom and XGLM which have different training data but similar sizes. Due to GPU memory limitations we did not study larger models and it is not clear whether findings generalise to even larger models.
- Although the TED talks dataset is a good overall testbed because it covers many topics and combines formal language and informal text, we did not quantify whether coherence is more likely to affect formal or informal language and might be studied with other datasets.
- This paper focuses heavily on MT as a complex generative task to study coherence of context, and it is not immediately clear whether the findings would also generalise to other longer-context generation tasks such as document summarization or how this would affect simple classification.

9 Ethical Concerns

Large LMs are known to hallucinate text content, potentially produce toxic speech or misinformation. While we did not observe this frequently in our experiments, we did not quantify the extent of this across various methods.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Human Language Technologies 2007: The Conference of the North American*

Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 436–443.

John Flowerdew and Michaela Mahlberg. 2009. *Lexical cohesion and corpus linguistics*, volume 17. John Benjamins Publishing.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Andreas Krause and Carlos Guestrin. 2008. Beyond convexity: Submodularity in machine learning. *ICML Tutorials*.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A Hearst. 2021. Can transformer models measure coherence in text? re-thinking the shuffle test. *arXiv preprint arXiv:2107.03448*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. *arXiv preprint arXiv:1809.00388*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

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

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. 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.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Yuan Wang and Minghe Guo. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2):460.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

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

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

A Resources

Software

- Implementation of Nearest Neighbor Retrieval with FAISS library (Johnson et al., 2019)
- Huggingface library was used for LLMs, model weights and calculation of perplexity.

Hardware

- All experiments can be run with a single NVIDIA-TITAN RTX GPU (24GB).

Datasets All datasets for the experiments are open-source.

B Local Coherence (nprompts=15)

Ablation experiments for Section 5.3 prompt set size of 15 shown in Table 7. The same trend is observed for 5 and 15 prompts.

C Domain vs Perplexity

We report the perplexity of the source sentences when randomly sampling and testing from different domains. Although there is no widely accepted measure of *general coherence*, we can formulate this with respect to the particular model being studied. We consider the model's conditional perplexity of the test sentence given the context. Perplexity is a widely used measure of suprisal in text and has also been used as a measure in topic coherence (reference). Concurrent work by Gonen et al. (2022) argue that total perplexity of the input sequence is related to In-context performance. We report the conditional perplexity from sampling and testing in different domains for GPTNeo2.7B in Table 8 and Bloom3B in Table 9. We did not report XGLM2.9B because the model log likelihood is very poorly calibrated.

D Local Coherence Shuffle Effects

BLEU scores for comparing `window` with other baselines, accompanying appendix section to Section 6.2 which reports BLEU scores for GPT-Neo2.7B `en`→`fr`. we find that results generalise across several models and languages that we further investigated.

	In/outdoc	GPTNeo2.7B(BLEU)			Bloom3B(BLEU)			XGLM2.9B(BLEU)		
		en-fr	en-pt	en-de	en-fr	en-pt	en-de	en-fr	en-pt	en-de
`random`	out	27.2	27.3	16.9	35.3	35.4	8.0	29.2	31.2	20.7
`nn`	out	27.3	28.2	17.1	35.6	35.9	9.1	30.0	32.0	21.6
`bm25`	out	27.9	29.0	17.4	36.1	36.4	**10.8**	31.2	33.0	22.2
`bm25-s`	out	27.7	29.1	17.3	35.2	36.0	9.1	29.8	32.0	21.6
`random`	within	28.1	29.2	17.6	36.8	37.3	8.9	30.9	33.3	22.3
`window`	within	**28.9**	**29.8**	**18.2**	**37.8**	**38.1**	9.6	**31.7**	**34.4**	**23.0**

Table 7: BLEU score comparison of similarity-based retrieval methods from out of document, and moving window (`window`) from within the document. 15 prompt examples used.

Prompt / Test	FLORES	MED	MTNT	TED
FLORES	**21.3**	24.5	54.9	25.5
MED	24.1	**16.2**	62.4	27.0
MTNT	25.4	26.9	**40.8**	23.3
TED	24.0	24.9	52.2	**19.6**

Table 8: Source sentence perplexity conditioned on prompts randomly sampled from the domain computed with GPT-Neo2.7B. Lower perplexity indicates greater coherence.

Prompt / Test	FLORES	MED	MTNT	TED
FLORES	**21.5**	23.4	61.8	28.5
MED	25.1	**16.3**	74.8	33.4
MTNT	25.2	25.7	**47.2**	26.1
TED	24.1	23.5	60.1	**22.1**

Table 9: Source sentence perplexity conditioned on prompts randomly sampled from the domain computed with Bloom3B. Lower perplexity indicates greater coherence.

	GPTNeo (en-pt)	GPTNeo (en-de)	XGLM (en-fr)	Bloom (en-fr)
`static`	27.1	16.8	27.7	34.9
`random`	27.3	17.3	26.6	35.9
`window`	28.3	17.9	28.5	36.9
`shuffle`	28.5	17.9	28.7	36.9

Table 10: BLEU scores for different ordering effects within document. All retrieval methods are within document.