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

**Suzanna Sia**                                                              ssia1@jh.edu
**Kevin Duh**                                                    kevinduh@cs.jhu.edu
Johns Hopkins University

**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. Code for this paper is available at `https://github.com/suzyahyah/icl_coherence_mt`.

## 1 Introduction

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). 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.

Previous work assumes that the role of the prompt context is to allow the model to "learn by examples". This has led to formulating the task of prompt selection as 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 use a heuristic version optimizing for word coverage.

| 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.

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 models 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

---

[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.

a test sentence. A similar working definition of coherence has been used in discrimination tasks that require a model to identify the right order of (shuffled) sentences (Elsner et al., 2007; Barzilay and Lapata, 2008; Laban et al., 2021).

## 3 Factors which affect In-context MT

We outline several factors studied in this paper related to example selection for in-context MT in Figure 1. 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.
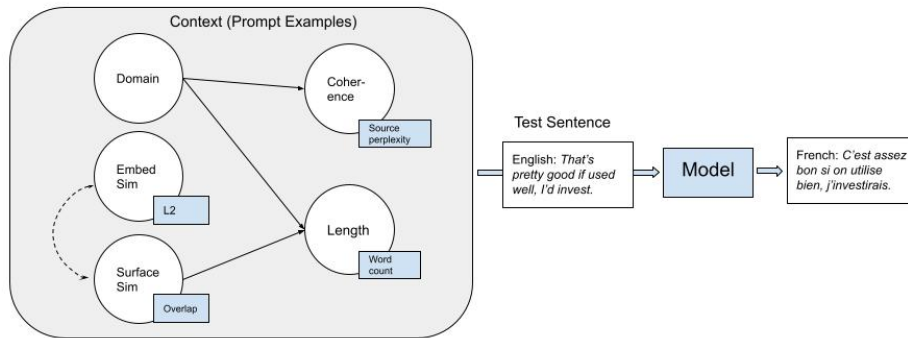


Figure 1: Factors identified and studied in this paper. Each domain has different length distributions (Section 5.2). Surface similarity and embedding similarity are associated (Table 4). Surface similarity selection also results in longer sentences (Section 5.4) Rectangle boxes next to the node are measures of these factors.

### 3.1 Length (Translation Budget)

One previously overlooked factor is the length (number of words) of 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 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 $f : 2^U \to R_+$ 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 MT.

## 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).

---

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

---
**Algorithm 1:** Generalised greedy (submodular) algorithm with length budget
---
1 **Input:** (Submodular) function $f : 2^U \to R_+$, cost function $m$, budget $b$, finite prompt bank $U$
2 **Output:** $X_k$ where $k$ is the number of iterations/prompts.
3 Set $X_0 \leftarrow; i \leftarrow 0;$
4 **while** $m(X_i) < b$ **do**
5     $u_i = \text{argmax}_{u \in U \setminus X_i} f(\{u\} \mid X_i)$
6     $X_{i+1} \leftarrow X_i \cup u_i;$
7     $i \leftarrow i + 1$
---

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". We also experimented with OPT2.7B, but find that its incontext MT abilities were nearly twice as poor as GPTNeo2.7B. 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.

## 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

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

| Prompt / Test | GPTNeo2.7B | | | | Bloom3B | | | | XGLM2.9B | | | |
| | 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.

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.

# 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 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**

- Multilingual GPT models namely Bloom and XGLM appear to be doing domain adaptation, as sampling and testing within the same domain (e.g., sample from MED test with MED) mostly results in the highest performance column-wise.

- 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).
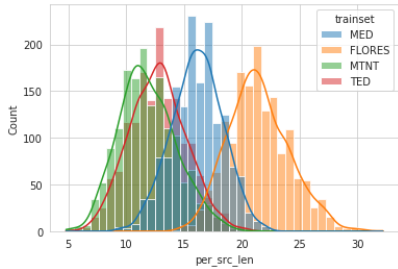
178

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 | ⇓18.5 | ↓20.8 | ⇓22.5 |
| MTNT | ⇓23.2 | ↓18.3 | ⇓21.9 | ↓23.5 |
| TED | ⇓21.7 | ↓17.6 | ⇓20.1 | ⇓22.3 |
| 5-10 words long sentences; GPTNeo 2.7B | | | | |
| FLORES | 24.2↓ | 19.6 | 22.7↓ | 24.3↓ |
| MED | 22.9 | 19.3 | 21.1 | 22.8↓ |
| MTNT | 24.0↑ | 18.9↑ | 22.5 | 24.3⇑ |
| TED | 23.8⇑ | 19.0↑ | 22.9⇑ | 23.8 |
| 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.

**Results and Discussion**

- 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 performance 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

---

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

| | In/outdoc | GPTNeo2.7B(BLEU) | | | Bloom3B(BLEU) | | | XGLM2.9B(BLEU) | | | L2 | coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 | 24.9 | 26.7 | 18.9 | 1.35 | 0.31 |
| nn | out | 26.8 | 26.9 | 16.9 | 35.1 | 35.1 | 8.2 | 25.4 | 26.6 | 18.3 | 0.98 | 0.49 |
| BM25 | out | 27.1 | 27.4 | 17.3 | 35.1 | 35.3 | **9.4** | 25.9 | 27.0 | 18.4 | 1.21 | 0.75 |
| BM25-s | out | 27.2 | 27.5 | 17.4 | 34.8 | 34.9 | 9.1 | 25.4 | 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 | **26.7** | **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.

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. Our results are consistent with concurrent work by Karpinska and Iyyer (2023) who show that translating an entire document is more effective than sentence by sentence 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).

  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.
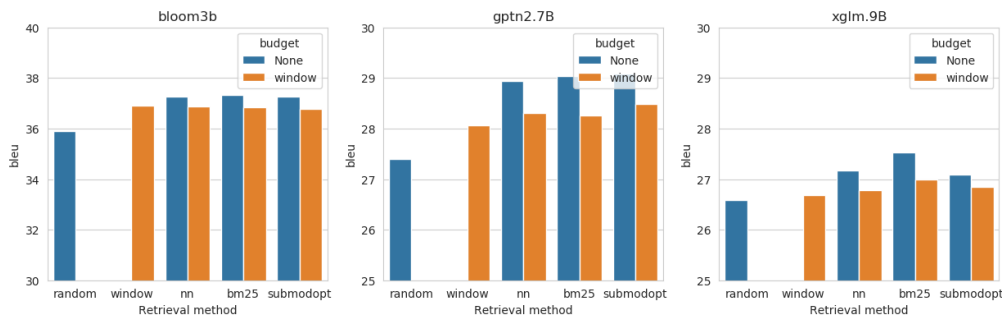
Figure 3: Comparison of Retrieval methods controlling for length budget: No budget or same budget as moving window. `random` is sampled within the document.

**Results and Discussion**

- 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 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.
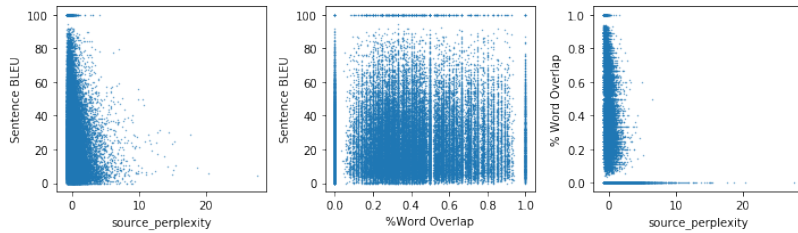
181

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

| 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.

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. The ordering of the document does affect source perplexity, with perplexity increasing from $11.1 \rightarrow$ 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.

## 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 found that 2 out of 3 models are 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 window 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

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.

# References

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2022). In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R. L., Koehn, P., Leiva, L. A., 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.

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

Bawden, R., Bretonnel Cohen, K., Grozea, C., Jimeno Yepes, A., Kittner, M., Krallinger, M., Mah, N., Neveol, A., Neves, M., Soares, F., Siu, A., Verspoor, K., and Vicente Navarro, M. (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.

Black, S., Leo, G., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carbonell, J. and Goldstein, J. (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.

Duh, K. (2018). The multitarget ted talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Elsner, M., Austerweil, J., and Charniak, E. (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.

Flowerdew, J. and Mahlberg, M. (2009). *Lexical cohesion and corpus linguistics*, volume 17. John Benjamins Publishing.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gonen, H., Iyer, S., Blevins, T., Smith, N. A., and Zettlemoyer, L. (2022). Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Karpinska, M. and Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Krause, A. and Guestrin, C. (2008). Beyond convexity: Submodularity in machine learning. *ICML Tutorials*.

Laban, P., Dai, L., Bandarkar, L., and Hearst, M. A. (2021). Can transformer models measure coherence in text? re-thinking the shuffle test. *arXiv preprint arXiv:2107.03448*.

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., del Moral, A. V., Le Scao, T., Von Werra, L., Mou, C., Ponferrada, E. G., Nguyen, H., 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*.

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

Lin, H. and Bilmes, J. (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.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2021). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2021). What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. *arXiv preprint arXiv:1809.00388*.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (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.

Philip, J., Berard, A., Gallé, M., and Besacier, L. (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.

Post, M. (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.

Reif, E., Ippolito, D., Yuan, A., Coenen, A., Callison-Burch, C., and Wei, J. (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.

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

Rubin, O., Herzig, J., and Berant, J. (2021). Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

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

Webson, A. and Pavlick, E. (2021). Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (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.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.