**Milestone 1**

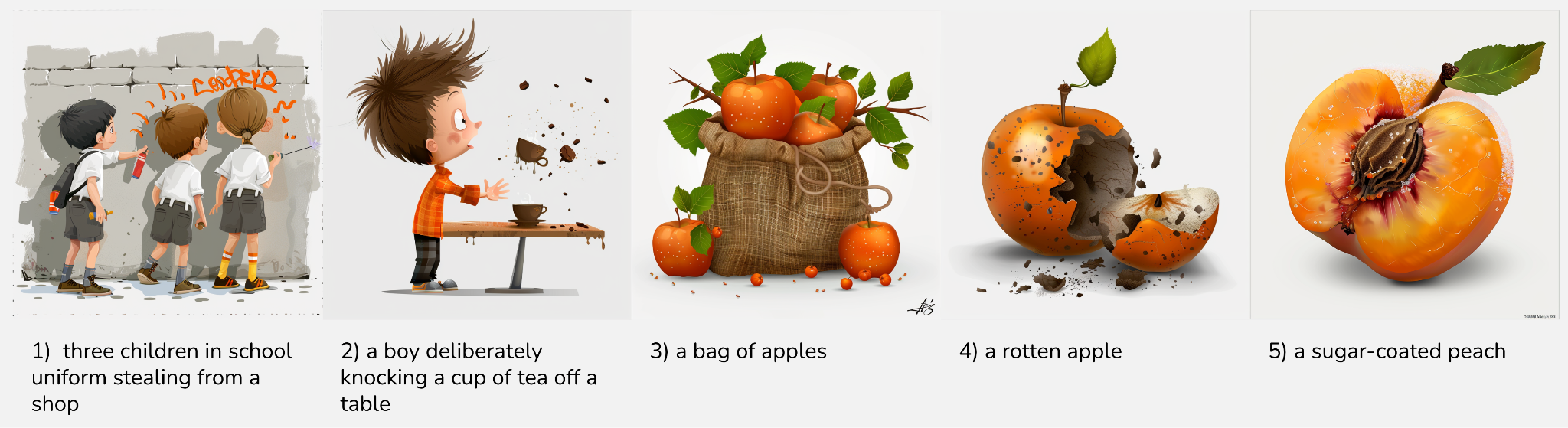
**The Problem Definition (Advancing Multimodal Idiomaticity Representation):**

Idioms are a class of multi-word expressions (MWE), which pose a challenge for current state-of-the-art models because their meanings are often different from the individual words that compose them.

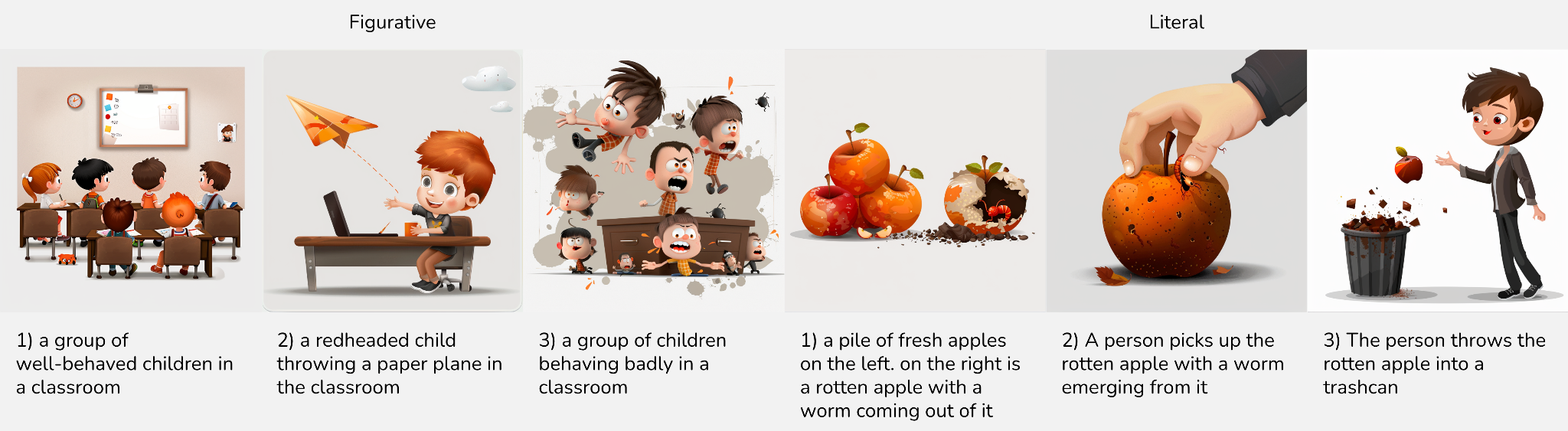
For instance, "piece of cake." doesn't typically refer to an actual slice of cake; instead, it refers to something that is easy to do. These expressions may also generate confusion between the literal, surface meaning arising from their component words and the idiomatic meaning. These, among other characteristics, make them a valuable testing ground for examining how NLP models capture meaning.

We present two subtasks:

* Subtask A - Static Image: The model will be presented with a set of 5 images and a context sentence in which a particular potentially idiomatic nominal compound (NC) appears. The goal is to rank the images according to how well they represent the sense in which the NC is used in the given context sentence.



* Subtask B - Image Sequences (or Next Image Prediction): The model will be given a target expression and an image sequence from which the last of 3 images has been removed, and the objective will be to select the best fill from a set of candidate images. The NC sense being depicted (idiomatic or literal) will not be given, and this label should also be output.



**Evaluation metrics:**

* Accuracy
* F1 Score

**First Research Paper: Evaluating the use of Large Language Models for Idiomaticity Detection**

**Experiment definition**

The paper evaluates the performance of large language models, including GPT-3.5 and GPT-4, on idiomaticity detection tasks. It highlights the differences in model performance based on the nature of the tasks and the datasets used. The authors note that while larger models (e.g: GPT-4) generally perform better, they do not necessarily outperform smaller, fine-tuned models specifically designed for idiomaticity detection (e.g: Flan-T5-XXL). The paper also discusses the impact of prompt engineering and few-shot prompting on model performance, particularly for less common languages like Galician. The findings suggest that while high-scale models can achieve competitive results, there are still significant challenges in accurately detecting idiomatic expressions, especially when the models are not fine-tuned for specific tasks.

**Methodology, Dataset, and Implementation Details**

To compare the performance of LLMs for idiomaticity detection, the study focuses on three datasets: FLUTE, SemEval 2022 Task 2a, and MAGPIE.

1. **Models Selected**:
   * SaaS models: GPT-3.5-turbo, GPT-4, GPT-4-turbo, and Gemini Pro.
   * Local models: Llama-2 (7B and 13B), Mistral-7B, and Phi-2.
   * Multilingual models: Flan-T5 variants (Small, Base, Large, XL, and XXL).
2. **Datasets**:
   * **FLUTE**: Evaluates idiom understanding as a natural language inference task.
   * **SemEval 2022 Task 2a**: A binary classification task determining if noun compounds are idiomatic or literal.
   * **MAGPIE**: Contains annotations for potentially idiomatic expressions, labeled as idiomatic, literal, or other.
3. **Experiment Setup**:
   * Prompts were designed for zero-shot and few-shot settings. For example, the SemEval prompt required models to disambiguate idiomatic and literal contexts with a binary classification response (i for idiomatic, l for literal).
   * SaaS models were evaluated using APIs, while quantized versions of local models were run on consumer-level hardware.
4. **Metrics**: Performance was assessed using macro-average F1 scores for consistency across all datasets.

**Findings and Results**

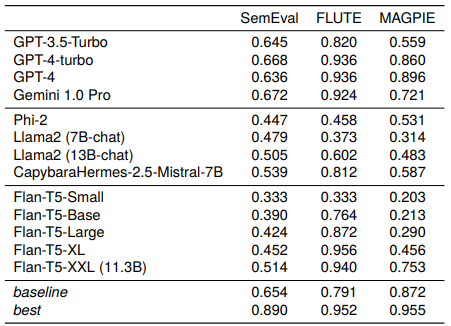


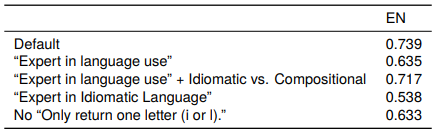
Table 3: Main results of our models across the three idiomaticity datasets. All results presented are macro-average F1 scores over the two classes. Baseline results are taken from Madabushi et al. (2021), Chakrabarty et al. (2022) and Zeng and Bhat (2021). ‘Best’ results (in all cases using models fine-tuned on the task training data) are taken from Chu et al. (2022), Bigoulaeva et al. (2022) and Zeng and Bhat (2021). For SemEval, the ‘zero-shot’ setting is reported.

Table 4: Results (macro F1) on the English test set of SemEval with GPT-3.5-turbo using prompt engineering

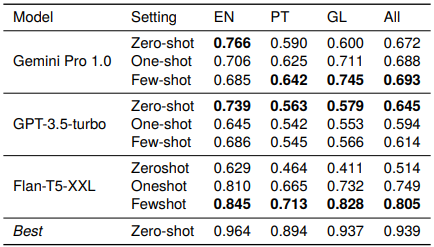


Table 6: Results on SemEval using few-shot prompting.

**Second Research Paper: Idiom Detection with Conversational Large Language Models**

**Experiment definition**

Our experiment aimed to evaluate the ability of selected LLMs to detect idiomatic expressions in sentences. This "idiom detection" task involved binary classification, predicting whether a sentence was "idiomatic" (positive class) or "non-idiomatic" (negative class). The goal was to assess if LLMs could accurately distinguish between figurative and literal meanings based on sentence formulation. Given that the pre-training data likely contained idiomatic expressions more frequently than literal ones, the models might be biased towards attributing figurative meanings based on probability distribution.

**Methodology, Dataset, and Implementation Details**

To ensure fair comparability among models, we focused on three LLMs with common characteristics: transformer-based architecture, approximately 7 billion parameters, open-source availability, and fine-tuning for dialogue. We preferred open-source models for transparency, reproducibility, and cost-effectiveness. The selection of smaller models allowed for resource-efficient experiments on local machines without GPUs. The chosen instruction fine-tuned conversational models simulate real-world chatbot interactions. The models assessed were:

* Llama-2-7b-chat (Touvron et al., 2023)
* Mistral-7b-Instruct (Jiang et al., 2024).
* Vicuna-7b (Zheng et al., 2023).

We established two evaluation levels: a fully automatic evaluation and a comprehensive manual evaluation with error analysis.

### Dataset Creation Process

1. **Idiom List Creation:** A curated list of idioms was manually compiled from various online sources, ensuring a diverse range of syntactic structures. The list includes fixed expressions (e.g., "Nothing to write home about") and more flexible constructions (e.g., "To blow your own trumpet"). A total of 93 idioms were selected, each with a unique identifier and source.
2. **Idiomatic Sentence Crafting:** New sentences were crafted for each idiom to avoid data contamination. A small-scale crowdsourcing effort involved eight native English speakers with high linguistic proficiency, who created 164 idiomatic sentences that reflect natural language use.
3. **Distractor Sentence Crafting:** To create the negative class, 86 distractor sentences were generated. These sentences contained words from idiomatic expressions used in a literal context, posing challenges for LLMs. Various strategies were employed, including recontextualizing idioms, altering syntactic roles, and using phonetically similar phrases.
4. **Sentence Proofreading and Final Layout:** Efforts were made to avoid multiple idiomatic expressions in a single sentence and to mitigate gender bias by using gender-neutral language where possible. Each sentence was assigned a unique identifier indicating whether it was idiomatic or a distractor, resulting in a final dataset of 250 sentences (164 idiomatic and 86 distractor).

**Findings and Results**

First level of evaluation

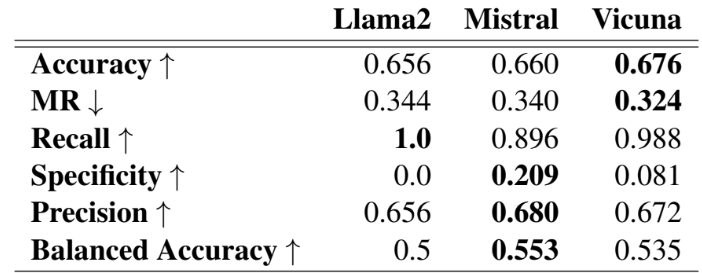
Accuracy, Misclassification Rate (MR), Recall, Specificity, Precision, and Balanced Accuracy. These metrics offer a general overview of the behaviour of the models and facilitate comparisons. In Table 1 we present the aggregated results. At a broad level, all three models fall within the same range of results, as they show close scores in terms of Accuracy and Misclassification Rate.

Table 1: Automatic metrics calculated for the three models. Numbers in bold indicate which model achieved the best result for each metric. For Misclassification Rate, lower values are indicative of better performance, as denoted by the downward arrow.

Second level of evaluation

In our study, we observed that in a certain number of cases the models, despite correctly classifying a sentence as idiomatic, did not detect the correct idiom and rather identified some other part of the sentence as idiomatic, such as a phrasal verb, a collocation, or a single word.

This observation underscores that general metrics are insufficient to conclusively demonstrate the capability of a LLM to detect an idiom in a sentence and motivated us to perform an additional verification step to validate the accuracy of true positive classifications. We calculated True Positive Consistency as the proportion of true positive predictions where the correct idiomatic expression was accurately identified as well. This additional score allowed us to validate whether the models response was grounded in the correct reason.

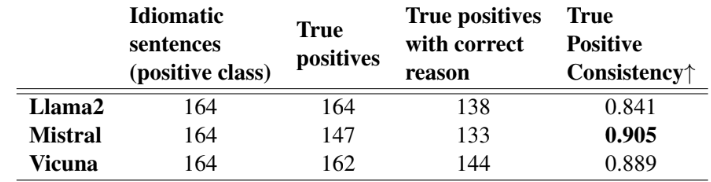
Table 2 displays True Positive Consistency values for the three analysed models. Mistral exhibits the best score, achieving a True Positive Consistency of 0.905, followed by Vicuna, and lastly Llama2.

Table 2: True Positive Consistency values per model.

**Third Research Paper: Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation**

**Experiment definition**

This paper investigates whether the non-compositionality of idioms is reflected in the mechanics of the dominant neural machine translation (NMT) model, Transformer, by analyzing the hidden states and attention patterns for models with English as the source language and one of seven European languages as the target language.

When the Transformer identifies the expression as idiomatic, the encoder processes idioms more strongly as single lexical units compared to literal expressions. This manifests in idioms' parts being grouped through attention and in reduced interaction between idioms and their context. In the decoder's cross-attention, figurative inputs result in reduced attention on source-side tokens. These results suggest that the Transformer's tendency to process idioms as compositional expressions contributes to literal translations of idioms.

After all, not all potentially idiomatic expressions (PIEs) are figurative – e.g. consider “When I kicked the bucket, it fell over”. Whether PIEs should receive a figurative or literal translation depends on the context.

**Methodology, Dataset, and Implementation Details**

This paper analyzes idiom processing for pre-trained NMT Transformer models for seven European languages (Dutch, German, Swedish, Danish, French, Italian, and Spanish) by comparing literal and figurative occurrences of PIEs. Large-scale analyses of idiom translations suffer from a lack of parallel corpora (datasets where idioms and their accurate translations are paired across languages).

The transformer contains encoder and decoder networks with six self-attention layers each and eight heads per attention mechanism. The models are pre-trained by Tiedemann and Thottingal with the Marian-MT framework on a collection of corpora (OPUS). We extract hidden states and attention patterns for sentences with PIEs. The analyses presented are detailed for Dutch, after which we explain how the results for the other languages compare to Dutch.

Parallel PIE corpora are rare, exist for a handful of languages only, and are limited in size. So, the paper uses the largest corpus of English PIEs to date and annotates the translations heuristically. The MAGPIE corpus contains 1756 English idioms from the Oxford Dictionary of English with 57k occurrences. MAGPIE contains identical PIE matches and morphological and syntactic variants, through the inclusion of common modifications of PIEs, such as passivization (“the beans were spilled”) and word insertions (“spill all the beans”). The paper uses 37k samples annotated as fully figurative or literal, for 1482 idioms that contain nouns, numerals, or adjectives that are colours (which we refer to as keywords). Because idioms show syntactic and morphological variability, the paper focuses mostly on the nouns. Verbs and their translation are harder to identify due to the variability. Moreover, idiom indexes are also typically organised based on the nominal constituents, instead of the verbs. Only the PIE and its sentential context are presented to the model. The paper distinguishes between PIEs and their context using the corpus’s word-level annotations.

The MAGPIE sentences are translated by the models with beam search and a beam size of five. The translations are labeled heuristically:

* Word-to-word translation: In the presence of a literal translation of at least one of the idiom’s keywords.
* Paraphrase translation: When a literally translated keyword is not present.

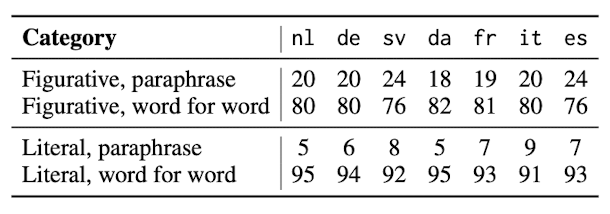
Table 1 summarises the distribution of these categories for all languages, for the subsets of figurative and literal examples from MAGPIE. The vast majority of literal PIEs indeed result in word-for-word translations. The subset of figurative samples results in more paraphrases, but ≥ 76% is still a word-for-word translation, dependent on the language. Although the statistics are similar across languages, there are differences in which examples are paraphrased. Figure 2 illustrates the agreement by computing the F1-score when using the predictions for figurative instances of one language as the target, and comparing them to predictions from another language. 

Table 1: Distribution of the heuristically assigned la- bels for translations of MAGPIE sentences in percent- ages, expressed within category (figurative / literal).

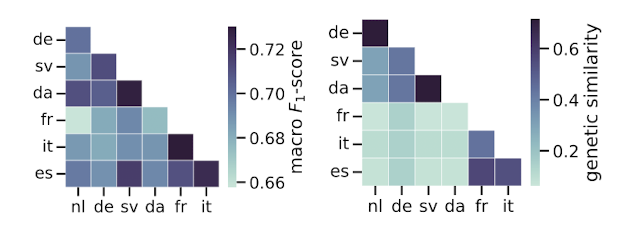


Figure 2: The macro-averaged F1-score of translation labels (paraphrase vs word for word) for figurative PIEs and languages’ genetic similarity visualised (Pearson’s r =0.61, p < 0.005).

To assess the quality of the heuristic method, one (near) native speaker per target language annotated 350 samples, where they were instructed to focus on one PIE keyword in the English sentence. Annotators were asked whether:

1. The English word was present in the translation (referred to as “copy”).
2. There was a literal translation for the word.
3. Neither of those options were suited, referred to as the “paraphrase”.

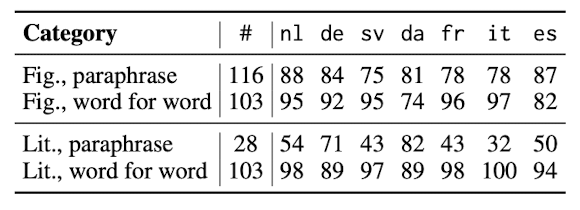
Due to the presence of cognates in the “copy” category, that category was merged with the “word for word” category after the annotation. Table 2 summarises the accuracies obtained. Of particular interest are samples that are figurative and paraphrased, since they represent the translations that are treated non-compositionally by the model, as well as instances that are literal and translated word for word, since they represent the compositional translations for non-idiomatic PIE occurrences. These categories have annotation accuracies of ≥ 75% and ≥ 89%, respectively. During preliminary analyses, an annotation study was conducted for Dutch by annotators from the crowd-sourcing platform Prolific. The annotators and the heuristic method agreed in 83% of the annotated examples, and for 77% of the samples, an average of 4 annotators agreed on the label unanimously.

Table 2: Survey statistics: the number of sentence pairs used (#), and the percentage of labels for which the annotator and the algorithm agreed per language.

**Findings and Results**

#### Encoder's Behavior

* Grouping of Idioms: For figurative idioms, the encoder groups components more strongly, treating them as a single unit.
* Reduced Contextual Interaction: Figurative idioms show reduced attention to surrounding context, reflecting standalone processing.

#### Decoder's Behavior

* Figurative idioms exhibit reduced reliance on source tokens and increased reliance on special tokens.
* This suggests the decoder generates translations more independently, contributing to paraphrased outputs but potentially undermining literal translations.

#### Probing Results

* Hidden Representations: Higher encoder layers capture figurativeness more distinctly.
* Amnesic Probing: Manipulating hidden representations causes shifts toward literal translations, establishing a causal link between encoded features and translation outcomes.

**References:**

1. Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. [Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection](https://aclanthology.org/2024.mwe-1.22). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
2. Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. [A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models](https://aclanthology.org/2024.figlang-1.5). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
3. Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation](https://aclanthology.org/2022.acl-long.252). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

**Group Members:**

* Ahmad Arif - 444002984
* Musab Iskandar - 444003841
* Yousef Koshak - 444000774