CausalSense: Leveraging Common Sense Knowledge and LLMs for Joint Event Extraction and Relation Classification

Youssra Rebboud, Pasquale Lisena, Raphael Troncy
EURECOM, Sophia Antipolis, France
youssra.rebboud@eurecom.fr, pasquale.lisena@eurecom.fr, raphael.troncy@eurecom.fr

Abstract

Event Relation Extraction (ERE) aims to identify and classify semantic relationships between events expressed in texts. While existing work has mainly addressed temporal or simple causal links, fine-grained causal relations such as enable, prevent, and intend remain insufficiently explored, partly due to limited and imbalanced labeled datasets. We present a novel framework that leverages large language models (LLMs) and common-sense knowledge to jointly perform event extraction and relation classification. Our contribution includes (1) the creation of the CausalSense large-scale dataset containing more than 500k sentences from news data and common sense knowledge extracted from ATOMIC, and enriched synthetically; and (2) the evaluation of multiple architectures, including transformer-based models and end-to-end multitask systems for extracting fine-grained causal relationships. Experimental results show that our best-performing model achieves a 32.3% improvement in average F1-score over the current state of the art. The integration of common sense knowledge substantially enhances fine-grained causal relation detection. The CausalSense dataset, along with our code and models, is released as open source to support future research on causal event relationship extraction.

Keywords: Event Relation Extraction, Data Augmentation, Generative AI

1. Introduction

Events are unique occurrences involving one or more participants at a specific time and location, that can be described by specific text (Liu et al., 2020). The extraction of these events, along with the relationships between them, has been a longstanding focus of interest within research communities under the task named Event Relation Extraction (ERE). The existing literature has primarily studied temporal and sub-event relationships between events – such as in the TempEval initiative (UzZaman et al., 2013) and several related work (Wang et al., 2020; Huttunen et al., 2002) – with recent attention turning towards causality (Yang et al., 2021). However, relations between events can go beyond simple causality, encompassing more semantically precise and fine-grained forms of connection such as enabling, prevention and intention. These precise relations have received limited attention in research and produced limited results so far (Rebboud et al., 2023), yet they are crucial for applications requiring deep semantic understanding, such as news analysis and narrative generation (de Kok et al., 2024). This can be attributed to two factors: the lack of labeled datasets representing these fine-grained event relations, and the inherent complexity of understanding these relations from text.

In this work, we aim to fill this gap in proposing a dataset and a model that accurately extracts fine-grained event relations, such as direct causality, enablement, prevention and intention. This model can identify these event relations and detect the cases where no such relation exists, together with the corresponding text spans. As many causal relationships reflect everyday knowledge, we propose to leverage common-sense knowledge for training our model. We hypothesize that by teaching the model fundamental common sense relations – such as *if you touch fire, you get burned* or *vaccines prevent complications* –, it can develop a foundational understanding of causality. This base knowledge can then be fine-tuned to grasp more complex causal patterns, such as those found in news data. Finally, our method makes use of Large Language Models (LLM) for data augmentation to further enrich the CausalSense dataset.

The contributions of this work are as follows:

1. We create CausalSense, a **dataset of over 500,000 sentences** annotated with **five fine-grained event relation types**, addressing the limitations of existing datasets in terms of size and class balance.

2. We provide a **model trained on this dataset**, capable of extracting fine-grained event relations from text.

3. We thoroughly **evaluate different approaches** (pre-trained language models, sequence-to-sequence models, and LLMs) on this dataset.

Our implementation, experimental environment and dataset are available at `https://github.com/ANR-kFLOW/Relation_extraction`.

The remainder of this paper is organized as follows. After providing an overview of the related work in Section 2, we detail our approach for creating CausalSense in Section 3. We describe our

ERE approach in Section 4. We analyze and discuss the results in Section 5. Finally, we conclude and outline future work in Section 6.

2. Related Work

Various techniques have been employed to extract event relations from texts, including supervised, unsupervised, semi-supervised, and distant supervision methods. These approaches primarily focus on extracting specific event relations, such as causality, temporality, and coreference (Liu et al., 2020; Li et al., 2021). Auto-regressive sequence-to-sequence models, such as REBEL, classify event relations by extracting triples *(subject, relation, object)* from text (Huguet Cabot and Navigli, 2021).

Large Language Models (LLMs) have also shown promise in this domain. For example, the Flan-T5 model combined with chain-of-thought prompting and few-shot learning using GPT-3.5 outperformed the REBEL baseline for relation extraction on the CoNLL04 dataset (Wadhwa et al., 2023). Additionally, LLMs have been applied to causal relationship extraction from tabular data (Liu et al., 2024). Recent advances have used pre-trained language models (PLMs) to improve event relation extraction. Schrader et al. (2023) have combined data augmentation using GPT-3.5 with a RoBERTa-based multi-layer tagging approach to identify multiple causal relations in a single sentence. This method achieved top performance in the Event Causality Identification Shared Task 3 at CASE 2023 (Tan et al., 2022). ADELIE also uses LLMs to perform information extraction tasks, including relation detection and classification, by relying on a large-scale instruction-tuning dataset (Qi et al., 2024). In contrast, our methodology is grounded in formal definitions and curated examples, in alignment with domain-specific ontological frameworks. Hu et al. (2025) reformulate ERE as a question-answering task, prompting LLMs to determine if and how events are related. However, the relation types they investigate (temporal, causal, subevent, and coreference) only partially overlap with the relations studied in this work. Similarly, LearnDA, a framework designed for data augmentation in ERE, focuses exclusively on the general causal relationship (Zuo et al., 2021).

A key limitation in this area is the scarcity of labeled datasets tailored to fine-grained event relations beyond the widely represented temporal (UzZaman et al., 2013; Caselli and Vossen, 2017), mereological (Gottschalk and Demidova, 2019) and causal relations (Mirza et al., 2014). The Facts and Events Relationship Ontology (FARO) (Rebboud et al., 2022) addresses this issue by defining a structured schema that includes causality, intention, enabling, and prevention (Rebboud et al., 2022). It also introduces a first, albeit very small (663 English sentences), dataset annotated with such relations. However, this dataset is limited in size and it is imbalanced across the classes, hindering robust model training. A subsequent version (Rebboud et al., 2023) adds synthetic examples generated via GPT-3.5. While this approach achieved an average F1-score of 0.61 across relation detection and classification, performance on event extraction remained low, with an F1-score of approximately 0.33 Rebboud et al. (2023). Notably, the evaluation excluded the negative class (i.e. sentences with none of these relations), leaving the relation detection task only partially assessed.

3. CausalSense

To create a robust dataset for training an ERE system, we reconcile and augment existing fine-grained causality datasets. Our key idea is to incorporate common sense knowledge graphs and to leverage large language models for augmenting and balancing the data. Our hypothesis is that integrating news data with common sense knowledge will provide complementary information, thereby enhancing the effectiveness of supervised ERE approaches.

3.1. News Dataset

Our starting point is the Rebboud et al. (2023) dataset, for which, we used GPT 3.5 to augment a news dataset by introducing synthetically generated sentences through a few-shot prompting technique. To the pre-existing 663 English sentences, this generated data adds 1,228 new sentences to better cover the less represented relations, i.e. enable, prevent and intend.

The resulting dataset contains therefore a total of 1,891 sentences annotated with the type of relation that they contain, in addition to the textual spans representing the relation trigger, the subject and the object of the relation. Our experiments provided initial evidence that training models with synthetic data can enhance performance. Specifically, relation classification models showed improvements, increasing the average F1 score by 50%.

However, some limitations have also been observed:

- The dataset is still largely imbalanced, with the less represented classes (*no relation* and *direct cause*) having around a third of the examples of the most represented ones;

- The performance in the event extraction task is still limited, with an F1 score increasing from 0.12 to 0.33;

- We observed a potential data leakage issue, as approximately 21% of the test set shares over 90% semantic similarity with a small subset of training samples. This overlap may be explained by the use of test set seeds as illustrative examples during the LLM-driven data augmentation process, which could have unintentionally introduced similarities between the training and test splits.

Since this enhanced dataset is available, we opted to use it as our starting point, while resolving its main limitations: data overlap and class imbalance. To cope with the under-representation of the *direct cause* class, we explored additional pre-existing datasets containing causal relations. Specifically, we incorporated causal examples from the Causal News Corpus (CNC) (Tan et al., 2022) – introduced in (Tan et al., 2022) – that provides 3,417 annotated sentences, among which 1,811 contain cause-effect annotations, while the rest are labeled as non-causal.

After analyzing around 50 sentences from this dataset, we extract these 1,811 sentences as *direct causality* examples, that we later split into 1,710 examples in a training set and 101 in a test set.

To address the issue of data overlap, we removed sentences from the test set that had a high similarity with those in the training set, using Sentence-BERT embeddings (Reimers and Gurevych, 2019) – in particular `bert-base-uncased` (Devlin et al., 2019) – and cosine similarity (threshold of 90% similarity). On the other hand, we enhanced the test set by including additional realistic, real-world examples from the AVeriTeC (Schlichtkrull et al., 2023) dataset, described in (Schlichtkrull et al., 2023). After manually assessing a random subset of the dataset, we incorporated 216 new sentences containing the four studied causal relation types.

3.2. Adding Common-Sense Knowledge

To mitigate data sparsity and enhance model generalization beyond surface-level patterns, we augment our dataset with common-sense knowledge. Specifically, we extract causal event relations from ATOMIC (Sap et al., 2019), a large-scale common sense knowledge graph by Sap et al. (2019). In addition, since ATOMIC lacks coverage for certain types of relation such as *enabling* and *prevention*, we complement it by generating additional examples using LLMs.

This augmentation process involves:

- Prompting an LLM to generate diverse sentences for underrepresented relation types.

- Iteratively refining the outputs by re-prompting the LLM to enhance linguistic diversity and domain coverage.

The next subsections detail these two steps.

3.2.1. ATOMIC Common Sense Data

The *Atlas of Machine Commonsense (ATOMIC)* is a large-scale knowledge graph designed to enhance deep learning models' ability to perform *if-then* reasoning and reason about familiar events by leveraging crowd-sourced knowledge extraction (Sap et al., 2019). It contains more than 877k inferential knowledge tuples that describe common sense situations. Unlike traditional approaches that rely solely on taxonomic information from corpora, ATOMIC provides examples of everyday situations that are considered common sense.

Among those, we are interested in types that can be mapped to the FARO relations. We observe that some relations – *(o|x)Want/ xIntent* and *(o|x)Effect*[1] respectively overlap with the definitions of direct cause and intend. Consequently, we can safely include them in our dataset. We provide 3 examples of triples from ATOMIC mapped to FARO in Table 2.

3.2.2. Augmenting Common Sense Knowledge for Event Relation Extraction

ATOMIC lacks event relations such as *enabling* and *prevention*. Therefore, we expand the dataset by generating new examples using a LLM. We apply an iterative refinement process where the generated examples are fed back into the LLM for further augmentation, increasing variety, and ensuring broad domain coverage. The final prompt used is given in Figure 1 after a prompt engineering session.

To ensure high-quality generated data, we evaluate multiple LLMs. The selection process is based on a manual review of the initial outputs, assessing their coherence, diversity, and adherence to expected relation types. The best-performing model is then used for subsequent data generation and refinement.

We evaluated three open weight models that were available at that time: **Llama2** (Touvron et al., 2023), **Zephyr** (Tunstall et al., 2024), and **Truthful-DPO-TomGrc FusionNet**. Truthful-DPO, based on Mixtral, a sparse mixture of experts model (SMoE) developed by MistralAI, was the top-performing model on the Hugging Face leaderboard at the

[1] In ATOMIC, the prefix 'x' typically represents the person or entity which is the subject of an action. For instance, 'xIntent' signifies the intention of person X, the entity initiating the action. On the other hand, the prefix 'o' stands for 'others,' indicating the impact or perspective from the viewpoint of those affected by the action. For example, 'oEffect' denotes the effect of person X's actions on others, capturing the consequences or observed outcomes of their behavior.

Category	**Dataset**	**Total**	**Cause**	**Enable**	**Prevent**	**Intend**	**No-rel.**
News Data	Original Data (Rebboud et al., 2022)	663	268	100	81	42	172
	Synthetic Data (GPT3.5)	1,228	0	350	419	459	0
	CausalNews Corpus (CNC)	3,316	1,710	0	0	0	1,606
Common Sense	ATOMIC	315,173	82,242	0	0	146,588	86,943
	Synthetic Common Sense	205,884	0	65,485	53,456	0	86,943
Total		526,264	84,321	66,025	54,067	147,189	175,664
Combined dataset		6792	3520	814	948	944	566
Test dataset		632	351	89	52	40	100
	including AVeriTeC	216	133	46	26	11	0

Table 1: Statistics of the CausalSense dataset

Subject	**Relation**	**Object**	**Mapped relation**
PersonX looks before you leap	xIntent	to be cautious	intends-to-cause
PersonX looks towards PersonY	xWant	to greet PersonY	intends-to-cause
PersonX loses 15 pounds	xEffect	has more energy	causes

Table 2: Example triples from the ATOMIC dataset with FARO mappings

Objective:
Generate augmented sentences containing two events with a **RelationType** relationship.

Definition: *Relations and Events*

Task Instructions:

- Sentences must depict common-sense scenarios.
- Each sentence should include two events.
- The generated examples should adhere to the provided structure.
- Do not generate more than the requested number of sentences.

Output format: Generate examples with **RelationType** relations following the specified format.

Figure 1: Prompt structure for generating common sense examples for a given relation type (in the first iteration, we omit the examples part.)

time of development[2]. A manual review of the first 20 outputs shows that Zephyr and Truthful-DPO produced accurate annotations, whereas Llama2 struggled to correctly identify event spans, often misplacing the events within the sentence. For example, in the sentence "*Wearing a helmet prevents head injuries from occurring during extreme sports*", the model incorrectly considered *prevents head injuries* as the first event, while the actual causal event *wearing a helmet* was not properly identified, and no second event was tagged. Therefore Llama2 was excluded from further experiments.

To help the model identify cases where no causal relation exists between events, we introduce **negative samples**. This is achieved by restructuring the dataset into event triples and randomly swapping either the subject or the object in a way that invalidates the original relation. This approach ensures that the model learns to distinguish between valid and spurious event relations.

To mitigate potential error propagation from the generated data, we manually reviewed 100 generated examples for each relation type to identify recurring and systematic error patterns. This step was introduced specifically to reduce generation-related noise as much as possible, given that full manual correction of the dataset would be prohibitively expensive. Based on this analysis, we filtered out incorrect annotations by removing patterns that frequently led to misclassifications, such as sentences containing contrastive conjunctions or contradiction-related terms. For example, a LLM generated the following sentence for the "enable" relationship: "*Introducing salt into boiling water, it prevents...*"; the inclusion of the term "prevents" is a clear misclassification sign, so this sentence is filtered out.

This procedure does not constitute a full validation of the generated data. The overall reliability and effectiveness of the generated examples will ultimately be assessed through model performance on the final test set, rather than through this intermediate filtering step.

3.3. Combined Dataset

The final step consists in merging sentences from both the news and the common-sense datasets. We adopt a balanced sampling strategy, ensuring an equal number of examples per relation type from both sources. This composition allows us to

[2] https://huggingface.co/yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B

assess the impact of augmenting real-world textual data with generalizable causal patterns, while preserving the relevance of the target domain during training.

This final setup addresses two alternative experiments we performed but that proved to be underperforming, namely: i) merging the full common-sense dataset with news data (which caused a significant imbalance), and ii) sequential fine-tuning, first on common-sense data and then on news data. Table 1 summarizes the distribution of the examples in the final CausalSense dataset.

4. Event Relation Extraction

In this section, we describe our ERE approach that we apply on the generated dataset. Event relation extraction can be decomposed into three interconnected subtasks:

(1) **Relation Detection (RD)** involves identifying whether a causal relation exists or not between two events. This task can be framed as a binary classification problem;

(2) **Relation Classification (RC)** involves sequence classification, where sentences are classified into one of the target relations. The set of relations will be in our case: *cause*, *enable*, *prevent*, *intend*, or *no relation*;

(3) **Event Extraction (EE)** is the process of span detection, which precisely identifies the spans of text that represent the subject and the object of the relation, referred to as *event1* and *event2*.

To address these three tasks, we explore three strategies: 1. **Decomposition**: Train and test each subtask separately to see if breaking down the task reduces complexity and improves performance. 2. **Multitask Learning**: Use a single model to learn all three subtasks together, sharing representations across them. 3. **Prompting Large Models**: Leverage existing large models with few-shot prompting, with no training or fine-tuning required, just a few examples.

4.1. Decomposition

In this strategy, the three subtasks are addressed separately.

Relation Detection. This task is modeled as a binary sequence classification problem that aims to decide whether a particular sentence contains a fine-grained causal event relation or not. We categorize the sentences in our dataset into positive examples (class 1) if they contained relations such as cause, enable, prevent, or intend, and negative examples (class 0) otherwise. For this task, we trained a simple transformer-based binary classifier.

Relation Classification. This task constitutes a sequence classification problem with five distinct categories. For the relation classification, we trained a transformer-based model that receives a sentence as input, trained to classify the sentence into one of the five classes: *cause*, *enable*, *prevent*, *intend* or *no relation*, which are returned as output after a linear activation module.

Event Extraction. The detection of the events composing the relation is cast as a token classification problem. The spans are annotated following the BIO[3] tagging scheme (Ramshaw and Marcus, 1995). We also trained a transformer-based model to predict a BIO tag for each token. We provide below an example including a causal relations between the subject *prolonged drought* and the object *severe water shortage*:

"The($_{\text{O}}$) **prolonged**($_{\text{B-C}}$) **drought**($_{\text{I-C}}$) across($_{\text{O}}$) the($_{\text{O}}$) region($_{\text{O}}$) resulted($_{\text{O}}$) in($_{\text{O}}$) **severe**($_{\text{B-E}}$) **water**($_{\text{I-E}}$) **shortages**($_{\text{I-E}}$) and($_{\text{O}}$) crop($_{\text{O}}$) failures($_{\text{O}}$), leading($_{\text{O}}$) to($_{\text{O}}$) economic($_{\text{O}}$) hardship($_{\text{O}}$) for($_{\text{O}}$) local($_{\text{O}}$) farmers($_{\text{O}}$)."

For these three subtasks, we experiment with both BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021) base models. For relation detection and relation classification, we use the pre-trained version of the two models for sequence classification (Devlin et al., 2019; Zhuang et al., 2021). For the event extraction subtask, we use the models specifically pre-trained for the named entity recognition (NER) task.[4] These models are further fine-tuned on our dataset for fine-grained causality extraction.

During inference, we first use the relation detection model as a filter to exclude cases where no causal relation exists. Furthermore, we also train and test our model on detecting these cases (*no relation*) for the relation classification and event extraction subtasks. This allows us to evaluate their effectiveness in handling potential false positive leakage from the initial filter coming out of the relation detection model.

4.2. Multitask Learning

In this strategy, we adopt a unified multi-task learning framework in which a single model is trained to predict a complete causal triplet from an input

[3] BIO = *Beginning*, *Inside*, *Outside*

[4] https://huggingface.co/dslim/bert-base-NER,https://huggingface.co/51la5/roberta-large-NER

sentence. We implement and evaluate this strategy using two types of architecture: sequence-to-sequence (*seq2seq*) and sequence labeling (according to the definition provided by Zhao et al., 2024).

As a representative of a seq2seq architecture, we use REBEL (Huguet Cabot and Navigli, 2021), an auto-regressive model built on top of BART (Lewis et al., 2020), which has demonstrated strong performance in general relation extraction tasks. REBEL follows an encoder-decoder paradigm to convert raw input text into structured triplets in the form *(Subject, Relation, Object)*. This generation-based approach aligns naturally with the objective of refined causality extraction, enabling the model to learn context-aware relational structures directly from text.

Example input:

> "The government's swift action to impose a lockdown **prevents** the rapid spread of COVID-19 among the population."

Output:

> **<triplet>** impose a lockdown **<subj>** spread of COVID-19 **<obj>** prevent

While REBEL was not originally designed for fine-grained causal relation extraction, we fine-tune it on our dataset annotated with semantically precise causal relationships (Cause, Enable, Prevent, Intend, Not Cause) to assess its capacity to adapt to this more specialized task.

As a representative of sequence labeling approaches, we use a transformer-based architecture, which has shown strong performance in both generic causality and event extraction (Schrader et al., 2023; Kyriakakis et al., 2019). Our architecture extends RoBERTa with three task-specific heads:

1. **Relation Detection Head**: a binary classifier that identifies whether a causal relation exists in the input.

2. **Relation Classification Head**: a multi-class classifier that predicts the specific type of causal relation.

3. **Event Extraction Head**: a token-level BIO tagger that extracts cause and effect spans.

If no causal relation is detected, the model assigns the *no relation* label and tags all tokens with *O*. Otherwise, the model proceeds with relation classification and span extraction. This behavior is detailed in Algorithm 1.

Algorithm 1 Multi-Head RoBERTa pseudo-code for RE

```
Require: Input tokens $T$, batch size $B$, max seq length $L$, relation classes $R$, BIO labels $N$
Initialize: Encoder, classification heads (RD, RC, EE)
Initialize: Loss functions $\mathcal{L}_{RD}$, $\mathcal{L}_{RC}$, $\mathcal{L}_{EE}$
Encode input: $H \leftarrow$ Model$(T)$
Extract pooled output $P$ and sequence output $S$
Relation Detection:
$RD\_logits \leftarrow$ Linear$(P, 2)$
Compute loss: $\mathcal{L} \leftarrow \mathcal{L}_{RD}(RD\_logits, labels_{RD})$
Initialize Logits:
$RC\_logits \leftarrow \mathbf{0}_{(B,R)}$
$EE\_logits \leftarrow \mathbf{0}_{(B,L,N)}$
for $i \leftarrow 1$ to $B$ do
    if $\arg\max(RD\_logits[i]) = 1$ then
        $RC\_logits[i] \leftarrow$ Linear$(P[i], R)$
        $EE\_logits[i] \leftarrow$ Softmax(Linear$(S[i], N)$)
    else
        Assign $RC\_logits[i, 'no\_relation'] \leftarrow 1.0$
        Assign $EE\_logits[i, :, 0] \leftarrow 1.0$
    end if
end for
Compute additional losses:
$\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{RC}(RC\_logits, labels_{RC}) + \mathcal{L}_{EE}(EE\_logits, labels_{EE})$
Return: $\{\mathcal{L}, relation\_logits, type\_logits, span\_logits\}$
```

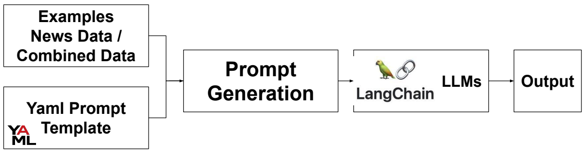

Figure 2: Workflow of ERE with LLMs

4.3. Prompting Large Models

LLMs have demonstrated strong performance across various NLP tasks, including language understanding, text classification, and information retrieval (Chang et al., 2024). Their effectiveness is driven by training on massive datasets and the application of advanced learning techniques such as self-supervised learning, fine-tuning, instruction tuning, reinforcement learning, and in-context learning, enabling them to generalize across a wide range of tasks (Zhai, 2024).

Given a sentence from our dataset, along with the definitions of each relation, we prompt the LLM to extract the subject, object, and relation from the text, as in Figure 3. The example part in the prompt is applied in the few-shot strategy, while it is omitted in the zero-shot scenario. The output of the LLMs is parsed to extract the subject, object, and relation from a given text. This process is repeated across the entire test set. The implementation architecture is represented in Figure 2 and makes use of

You are an expert in fine-grained causality extraction. Your task is to extract the subject, object, and relation from the given sentence. The relation must be one of the following: *cause*, *enable*, *prevent*, *intend*, or *no_relation* (if none of the refined causal relations apply).

Relation Definitions: *[The definitions of* ***cause*****,** ***intend*****,** ***enable*****, and** ***prevent*** *are based on the FARO ontology.]*

Task Instructions: Extract the **subject**, **object**, and **relation** for the following sentence.
Sentence: `"input_sentence"`

Output Format: `Subject: <extracted_subject>, Object: <extracted_object>, Relation: <extracted_relation>`

Important Guidelines:

- The relation must be one of: *cause*, *enable*, *prevent*, *intend*, or *no_relation*.
- Extract the **actual words** from the sentence for both **subject** and **object**.
- **DO NOT** use placeholders like ‘<subject>‘, ‘<object>‘, or ‘<relation>‘ in the output. Always provide extracted values from the input sentence.
- If the sentence does not contain any of the four refined causal relations, output: `Relation: no_relation`.

Examples: `examples`

Figure 3: Fine-grained causality extraction prompt that makes use of the relation definitions from FARO ontology.

the LangChain framework (Chase, 2022) for easy interaction with different LLMs.

In our experiment, we compare the performance of both a closed model (GPT-4o[5]) and the open weights model Zephyr-7B-beta-AWQ (Tunstall et al., 2024). The latter is a lighter variant of the Zephyr-7B-beta model[6] quantized with Activation-aware Weight Quantization (AWQ), which provides faster inference and requires less memory.

5. Results and Analysis

This section presents an analysis of the model performance across different datasets and strategies on each subtasks. The results are shown in Table 3.

5.1. Baseline

Our baseline builds upon the method proposed by Rebboud et al. (2023) that provides state-of-the-art results on the task of extracting fine-grained causal relationships from text. This approach was, however, originally evaluated on a slightly different test set – due to the data overlap issue mentioned in Section 3, solved in our dataset – and it did not address relation detection. For this reason we tested their model on our test set, in order to use it as a baseline.

The setup of this baseline uses BERT for all subtasks: BERT for sequence classification in relation classification, and BERT for token classification in event extraction. To extend this to the Relation Detection task, we adapted the labels by treating class 0 as the negative (*no relation*) class and all others as positive.

5.2. Discussion

Our best-performing model significantly outperforms the baseline (comparing with the best model for each task) achieving a 32.3% improvement in average F1-score for all tasks. Breaking down the performance, Relation Detection (RD) improves by 52%, Relation Classification (RC) by 24%, and Event Extraction (EE) by 32%, establishing our approach as the most effective to date for fine-grained causality extraction. End-to-end models outperform separate ones in 2/3 of the cases, highlighting the benefit of shared parameter learning.

The inclusion of common sense knowledge via the combined dataset further enhances performance, particularly for end-to-end models. Models trained on the combined set perform better or on par with their counterparts, with 50% cases showing clear improvements and only minor performance drops (at most 0.07%). To emphasize the impact of common sense integration, we underline in the results table all cases where performance improves. Specifically, we underline any instance where a model trained on the combined dataset outperforms its counterpart trained on the news dataset. These results provide strong evidence of the quality of the proposed dataset for the studied task.

Model-specific gains reflect this trend. REBEL trained on the combined dataset improves RD by 4%, RC by 10%, and EE by 11%. RoBERTa also benefits slightly in RC under end-to-end training (3%) in the separate task setting. In contrast, separate models mainly improve in RD, both BERT and RoBERTa achieved a 3% improvement, with little to no change in other subtasks, indicating that they may already capture implicit common-sense knowledge, limiting its added value.

LLMs show variable performance across tasks

[5] `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`

[6] `https://huggingface.co/HuggingFaceH4/Zephyr-7b-beta`

Dataset	Strategy	Model	Relation Detection			Relation Classification			Event Extraction			F1
			P	R	F1	P	R	F1	P	R	F1	
Baseline (Rebboud et al., 2023)	Separate	BERT	0.42	0.50	0.46	0.44	0.65	0.53	0.33	0.31	0.32	0.44
News	End-to-End	REBEL	0.86	0.42	0.56	0.75	0.60	0.65	0.58	0.65	0.59	0.60
		RoBERTa	0.98	0.98	**0.98**	0.79	0.71	0.74	0.19	0.20	0.20	0.64
	Separate	BERT	0.94	0.85	0.89	0.8	0.76	0.77	0.58	0.64	0.61	0.76
		RoBERTa	0.86	0.94	0.89	0.71	0.77	0.73	0.66	0.67	0.66	0.76
Combined	End-to-End	REBEL	0.72	0.51	0.60	0.77	0.74	0.75	0.74	0.68	**0.70**	0.68
		RoBERTa	0.99	0.96	**0.98**	0.75	0.83	**0.78**	0.19	0.20	0.20	0.65
	Separate	BERT	0.93	0.91	0.92	0.68	0.74	0.70	0.56	0.64	0.60	0.74
		RoBERTa	0.93	0.92	0.92	0.71	0.76	0.73	0.67	0.61	0.64	**0.763**
	LLM	GPT4 (0-shot)	0.18	0.85	0.29	0.54	0.37	0.33	0.37	0.23	0.23	0.29
		Zephyr (0-shot)	0.19	0.79	0.31	0.31	0.32	0.23	0.25	0.25	0.24	0.26
News	LLM	GPT4 (2-shot)	0.45	0.83	0.59	0.52	0.62	0.53	0.41	0.43	0.42	0.51
		GPT4 (4-shot)	0.40	0.97	0.57	0.55	0.64	0.54	0.46	0.44	0.45	0.52
		Zephyr (2-shot)	0.38	0.66	0.48	0.42	0.51	0.44	0.29	0.28	0.27	0.39
		Zephyr (4-shot)	0.17	0.98	0.29	0.28	0.22	0.10	0.23	0.21	0.20	0.19
Combined	LLM	GPT4 (2-shot)	0.34	0.88	0.49	0.49	0.58	0.46	0.36	0.35	0.35	0.43
		GPT4 (4-shot)	0.28	0.92	0.43	0.50	0.55	0.44	0.34	0.33	0.32	0.39
		Zephyr (2-shot)	0.22	0.94	0.36	0.41	0.38	0.30	0.29	0.29	0.29	0.31
		Zephyr (4-shot)	0.16	0.95	0.27	0.46	0.22	0.11	0.30	0.30	0.30	0.23

Table 3: Precision, Recall, and (macro) average F1-score across subtask, with the average F1-score on the combined task

and settings. GPT-4 consistently outperforms Zephyr and gains from few-shot prompting, reaching an F1-score of 0.52 (News, 4-shot) and 0.43 (Combined, 2-shot). LLMs do not benefit from common-sense examples, possibly due to their pretraining, and the reduced share of news examples in the combined set may further impact their results. The results consistently show a large gap between LLMs and fine-tuned PLMs. This calls for more extensive experiments using different prompting techniques, such as Chain-of-Thought.

5.3. Error Analysis

We conducted an in-depth analysis of the performance and error patterns across the ERE pipeline. While the overall performance for **relation detection** are optimal (up to 0.98) the **relation classification** remains more challenging, with a macro F1 score of 0.77. The most difficult relations are *prevent* (F1=0.64) and *enable* (F1=0.68), in contrast to *cause* and *no_relation*, which both reached an F1 score of 0.86.

The relation *enable* is inherently difficult to annotate and model, as it often involves subtle distinctions between a facilitating condition and a direct causal trigger. Human annotators frequently struggle to differentiate *enable* from *cause*, which may have lead the model to confusion. Similarly, *prevent* is often misclassified as *cause*, likely due to its underlying semantics (e.g. *A prevents B* resembles *A causes no B*) and the limited number of well-differentiated training examples. This is evident in the following example from our dataset, misinterpreted by our system as a causal relation, rather than a preventive intent: `The police have posted men in front of the office, [...] to prevent any retaliatory attack by RSS men.`

In **event extraction**, direct causality proved to be the most challenging relation, achieving an F1 score of 0.66. This is partly due to inconsistencies in annotation granularity, with certain examples being brief and localized, while others span full clauses[7], as evident from the following examples:

1. `<ARG1>The movement was catapulted into the headlines in early August</ARG1> when <ARG0>the semi-autonomous city [...] saw the first pro-independence rally</ARG0>`

2. `After a severe earthquake centered in [...] England, <ARG0>helped</ARG0> the region start to <ARG1>rebuild</ARG1>.`

6. Conclusion and Future Work

In this work, we studied the event relation extraction task with a focus on fine-grained causal relations that extend traditional causality. We introduce a model capable of accurately extracting these fine-grained causal relationships from text, achieving an average F1-score improvement of approximately 32.3% over the current state-of-the-art. Next to this, we released CausalSense, a novel dataset obtained by combining news data and common-sense

[7] In first approximation, sentences belonging to the first group were already present in the original dataset (Rebboud et al., 2023), while the second groups include sentences from the CNC subset. Despite these challenges, examples coming from CNC proved to be beneficial for training.

knowledge while being synthetically enriched with generated sentences.

Our experiments lead to the following key takeaway. The end-to-end approach generally proves more effective than the separate setup. While event extraction remains a challenging task, the inclusion of common sense knowledge yields noticeable improvements. LLMs show some potential in few-shot settings, but they still under-perform compared to fine-tuned PLMs. Our attempt to integrate common sense examples into LLM prompting led to a reduction in the number of news-based examples, without observable gains in performance, likely because LLMs already encode general common sense knowledge, making the added input redundant rather than complementary.

As future work, we plan to conduct ablation studies to isolate the impact of supervision noise and explore refined span alignment strategies to enhance performance on challenging relations such as *prevent* and *enable*. We would also like to measure the improvement provided by the introduction of different techniques for prompt-based ERE (Qi et al., 2024; Hu et al., 2025) and data augmentation (Zuo et al., 2021).

Furthermore, we plan to apply the capabilities of the developed system to a large corpus of news articles, enabling the extraction of semantically precise event relations. This process will facilitate the construction of a comprehensive knowledge graph representing events and their interconnections. Additionally, we aim to perform coreference resolution between our generated knowledge graph and other well-known knowledge bases such as EventKG (Gottschalk and Demidova, 2019) and Wikidata (Vrandečić and Krötzsch, 2014). We plan to employ link prediction and deletion techniques to refine the accuracy and reliability of the graph, with the hypothesis that it is possible to learn relation patterns and train a predicting algorithm from existing data. This can serve several scenarios, such as fact-checking in news.

Ethics Statement and Limitations

This work relies on datasets and language models that were either publicly released or generated under ethical and responsible data usage conditions, ensuring that no personally identifiable information or sensitive content was introduced. We acknowledge that generative models can reproduce or amplify social, cultural, or gender biases present in their training data. Care was taken to mitigate these effects through manual verification of generated samples and by filtering inconsistent or contradictory examples. Nevertheless, residual biases might persist and could influence downstream event-relation models.

While our models demonstrate strong performance, several limitations remain. First, relation classification suffers from confusion between semantically close categories such as *enable*, *cause*, and *prevent*, highlighting the intrinsic complexity of modeling fine-grained causal relations.

Second, while integrating additional data contributed to overall performance improvements, it also introduced inconsistencies due to broader and less precise annotation spans compared to the rest of our dataset. These mismatches, especially in the representation of causal constructs, may have hindered the model's ability to capture fine-grained distinctions. Future work could explore refined preprocessing or span alignment strategies to address this issue.

Finally, we have not yet explored alternative integration strategies such as contrastive learning, relation-specific adapters, or reinforcement-based rule alignment. These directions could potentially offer more robust generalization without relying on direct data augmentation. A final limitation is that our evaluation could always be extended to other models such as T5 (Raffel et al., 2019) and XLNet (Yang et al., 2019) for event relation extraction or new LLMs.

Previous attempts to integrate the full commonsense dataset with news data, including both direct merging and sequential fine-tuning, led to suboptimal results. As a result, we adopted a balanced setup with equal numbers of examples from each source. However, other integration strategies such as weighted sampling, or knowledge-aware curriculum learning remain unexplored and could be investigated in future work.

7. Acknowledgments

This work was supported by the French Public Investment Bank (Bpifrance) i-Demo program within the LettRAGraph project (Grant ID DOS0256163/00) and by the French National Research Agency (ANR) within the kFLOW project (Grant n°ANR-21-CE23-0028).

8. Bibliographical References

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan

Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).

Mike de Kok, Youssra Rebboud, Pasquale Lisena, Raphael Troncy, and Ilaria Tiddi. 2024. From Nodes to Narratives: A Knowledge Graph-based Storytelling Approach. In *Seventh International Workshop on Narrative Extraction from Texts (Text2Story), colocated with ECIR 2024*, Glasgow, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Gottschalk and Elena Demidova. 2019. EventKG – the hub of event knowledge on the web – and biographical timeline generation. *Semantic Web*, 10:1039–1070.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large Language Model-Based Event Relation Extraction with Rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496, Abu Dhabi, UAE. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and Joan Ginés i Ametllé. 2019. Transfer learning for causal sentence detection. In *18th BioNLP Workshop and Shared Task*, pages 292–297, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Junfei Liu, Shaotong Sun, and Fatemeh Nargesian. 2024. Causal Dataset Discovery with Large Language Models. In *Workshop on Human-In-the-Loop Data Analytics*, pages 1—-8. Association for Computing Machinery.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges. *AI Open*, 1:22–39.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating Causality in the TempEval-3 Corpus. In *EACL Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. ADELIE: Aligning Large Language Models on Information Extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Miami, Florida, USA. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*.

Youssra Rebboud, Pasquale Lisena, and Raphael Troncy. 2022. Beyond Causality: Representing Event Relations in Knowledge Graphs. In *Knowledge Engineering and Knowledge Management (EKAW)*, pages 121–135, Bolzano, Italy. Springer International Publishing.

Youssra Rebboud, Pasquale Lisena, and Raphaël Troncy. 2023. Prompt-based data augmentation for semantically-precise event relation classification. In *SEMMES 2023, Semantic Methods for Events and Stories, May 23-28, 2023, Heraklion, Greece*, Heraklion. CEUR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *33rd AAAI Conference on Artificial Intelligence*. AAAI Press.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Timo Pierre Schrader, Simon Razniewski, Lukas Lange, and Annemarie Friedrich. 2023. BoschAI @ Causal News Corpus 2023: Robust Cause-Effect Span Extraction using Multi-Layer Sequence Tagging and Data Augmentation. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 38–43, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct Distillation of LM Alignment. In *First Conference on Language Modeling*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *7^{th} International Workshop on Semantic Evaluation (SemEval)*, pages 1–9, Atlanta, USA. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting Relation Extraction in the era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64:1161 – 1186.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Neural Information Processing Systems*.

ChengXiang Zhai. 2024. Large language models and future of information retrieval: Opportunities and challenges. In *47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 481–490. Association for Computing Machinery.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11).

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen.

2021. LearnDA: Learnable Knowledge-Guided Data Augmentation for Event Causality Identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

9. Language Resource References

Chase, Harrison. 2022. *LangChain*.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. *BERT base model (uncased)*.

Rebboud, Youssra and Lisena, Pasquale and Troncy, Raphaël. 2022. *Facts and Events Relationship Ontology (FARO)*.

Rebboud, Youssra and Lisena, Pasquale and Troncy, Raphaël. 2023. *Fine-Grained Event Causality Dataset*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, et al. 2019. Atomic: An atlas of machine commonsense.

Schlichtkrull, Michael Sejr and Guo, Zhijiang and Vlachos, Andreas. 2023. *AVeriTeC Dataset*.

Tan, Fiona Anting and Hürriyetoğlu, Ali and Caselli, Tommaso and Oostdijk, Nelleke and others. 2022. *Causal News Corpus (CNC)*.

Hugo Touvron and Louis Martin and Kevin Stone and Peter Albert and Amjad Almahairi and Yasmine Babaei and Nikolay Bashlykov and Soumya Batra and Prajjwal Bhargava and Shruti Bhosale and Dan Bikel and Lukas Blecher and Cristian Canton Ferrer and Moya Chen and Guillem Cucurull and David Esiobu and Jude Fernandes and Jeremy Fu and Wenyin Fu and Brian Fuller and Cynthia Gao and Vedanuj Goswami and Naman Goyal and Anthony Hartshorn and Saghar Hosseini and Rui Hou and Hakan Inan and Marcin Kardas and Viktor Kerkez and Madian Khabsa and Isabel Kloumann and Artem Korenev and Punit Singh Koura and Marie-Anne Lachaux and Thibaut Lavril and Jenya Lee and Diana Liskovich and Yinghai Lu and Yuning Mao and Xavier Martinet and Todor Mihaylov and Pushkar Mishra and Igor Molybog and Yixin Nie and Andrew Poulton and Jeremy Reizenstein and Rashi Rungta and Kalyan Saladi and Alan Schelten and Ruan Silva and Eric Michael Smith and Ranjan Subramanian and Xiaoqing Ellen Tan and Binh Tang and Ross Taylor and Adina Williams and Jian Xiang Kuan and Puxin Xu and Zheng Yan and Iliyan Zarov and Yuchen Zhang and Angela Fan and Melanie Kambadur and Sharan Narang and Aurelien Rodriguez and Robert Stojnic and Sergey Edunov and Thomas Scialom. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*.

Lewis Tunstall and Edward Beeching and Nathan Lambert and Nazneen Rajani and Kashif Rasul and Younes Belkada and Shengyi Huang and Leandro von Werra and Clémentine Fourrier and Nathan Habib and Nathan Sarrazin and Omar Sanseviero and Alexander M. Rush and Thomas Wolf. 2024. *Zephyr 7B Beta - AWQ*.

Zhuang, Liu and Wayne, Lin and Ya, Shi and Jun, Zhao. 2021. *RoBERTa-large*.