# Using Large Language Models for Temporal Relation Extraction in Clinical Reports

## A Study on Patient Timeline Creation for Rare Diseases

Judith Jeyafreeda ANDREW, Mark VINCENT, Nicolas GARCELON, Anita BURGUN

## Abstract

This study evaluates Large Language Models for extracting temporal relations from clinical reports of pediatric rare disease patients. We aim to create accurate patient timelines by identifying relationships between phenotypes and temporal entities, potentially improving disease progression understanding and diagnosis.

## Introduction

- Focus: Temporal Relation Extraction for rare disease patient timelines
- Importance: Time and order of symptoms can differentiate related conditions
- Goal: Create accurate patient timelines from clinical reports

## Dataset

- 25 clinical reports (Necker Hospital)
- 706 annotated relations
- Pre-annotated phenotypes & temporal entities

## Relation Types

**BEGINS-AT**: Phenotype begins at time point

**ENDS-AT**: Phenotype ends at time point

**BEFORE**: Phenotype ends before time

**OVERLAP**: Share common time span

**CONTAINS**: Time contains phenotype span

**SIMULTANEOUS**: Same time span

**BEFORE-OVERLAP**: Occurs before and during time span

## Methods

### Models

- **Llama3**, **Gemma**, **Mistral**
- Locally deployed for privacy/security

### Approach

- **Multi-class classification**
- **Binary classification**
- 1000 tokens of context

## Limitations

- Limited dataset (25 clinical reports)
- French language data may influence results
- Computing constraints limited experiment scale

## Results

### Binary Classification F1 Scores

| Model | BEGINS-AT | ENDS-AT | CONTAINS | BEFORE |
|---|---|---|---|---|
| Llama | 0.56 | 0.55 | 0.18 | 0.48 |
| Mistral | **0.66** | **0.70** | 0.17 | **0.59** |
| Gemma | 0.62 | 0.57 | **0.40** | **0.59** |

| Model | OVERLAP | BEFORE-OVERLAP | SIMULTANEOUS |
|---|---|---|---|
| Llama | 0.03 | **0.67** | 0.10 |
| Mistral | 0.11 | 0.65 | 0.35 |
| Gemma | **0.37** | 0.66 | **0.37** |

### Multi-class vs Binary Performance

| Model | Multi-class Avg F1 | Binary Avg F1 | Improvement |
|---|---|---|---|
| Llama | 0.26 | 0.37 | +42% |
| Mistral | 0.32 | **0.46** | +44% |
| Gemma | 0.31 | 0.45 | +45% |

## Key Findings

- Binary classification prompt significantly outperforms multi-class approach
- Different LLMs excel at different relation types:
  - **Mistral**: Best for BEGINS-AT (0.66), ENDS-AT (0.70), BEFORE (0.59)
  - **Gemma**: Best for CONTAINS (0.40), OVERLAP (0.37), SIMULTANEOUS (0.37)
  - **Llama**: Best for BEFORE-OVERLAP (0.67)
- Complex relations (OVERLAP, CONTAINS, SIMULTANEOUS) remain challenging for all models
- Binary classification shows 42-45% improvement over multi-class approach

## Conclusion & Future Work

### Achievements

- Binary classification is more effective for relation extraction
- LLMs show promise for temporal relation extraction in healthcare
- Different models show strengths with different relation types

### Next Steps

- Improve detection of complex temporal relations
- Ensemble approach combining multiple models
- Expand dataset size for better generalization

***Contact Information:*** *Judith Jeyafreeda Andrew | Institut Imagine | judith.jeyafreeda@institutimagine.org*

Université Paris Cité

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

institut imagine
GUÉRIR LES MALADIES GÉNÉTIQUES