# UAlberta at SemEval-2026 Task 5: Disambiguating Stories via Task Decomposition

David Basil, Junhyeon Cho, Chirooth Girigowda, Guoqing Luo, Sahir Momin, Sevryn Robinson, Ning Shi, Grzegorz Kondrak

{dbasil1,gkondrak}@ualberta.ca

## Introduction

**Task 5:**

The goal is to predict the plausibility (on a 1-5 scale) that a word expresses a given meaning in varied contexts.

**We make note of two observations from prior work:**

1. Breaking tasks into pieces improves LLM robustness
2. LLMs struggle with numerical reasoning tasks

This motivates the design of our **Task Decomposition (TD)** system, which:

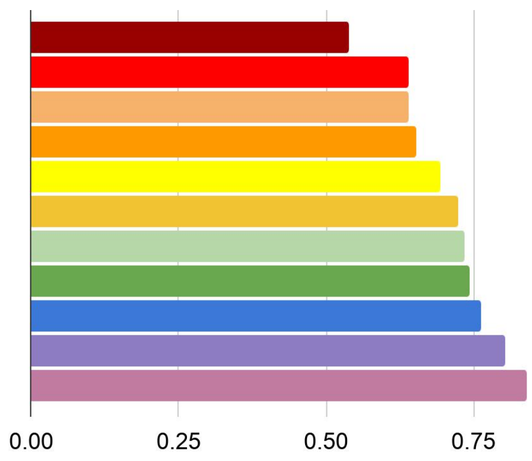
- Breaks Task 5 into *simpler*, *binary* subtasks for easier LLM inference.
- Combines the responses to these tasks into a final output using regression.

Our final system is a ridge regression **ensemble** of TD with other measures including **direct** LLM prompting, **WSD**, and **embedding**-based inference.

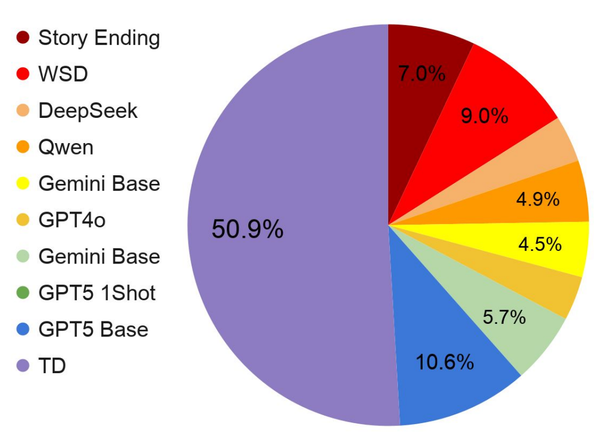
## Results and Analysis

We ensemble the described methods using **ridge regression**:

**Performance** (Spearman)




**Ensemble Weights**




- Ensemble **outperforms** all component systems.
- **TD** makes up the **majority** of the ensemble weight.

## Task Decomposition (TD)

- We prompt an LLM with four **binary** questions approximating Task 5.
- We also prompt the LLM to provide a plausibility score directly.
- These responses are encoded as **binary** and **numerical** features.
- We also include a binary indicator of whether the story has an **ending**.
- We train an XGBoost regressor on these features to predict a final plausibility score between 1 and 5.

| Subtasks: | Responses | Features | Output |
|---|---|---|---|
| **Which** of the two senses is more likely? | sense a | 0 | XGB Regressor → 4.25 |
| Is the sense **clear** or **ambiguous**? | clear | 1 | |
| Is the given sense **definitively** correct? | no | 0 | |
| Is **at least one** meaning plausible? | yes | 0 | |
| **Rate** the plausibility of the given sense | 4 | 4 | |
| (no ending in sample) | no | 0 | |

## Additional Methods

### Story Ending

- We fine-tune an **embedding** model to predict a plausibility score.
- We note that the data is generally designed so that the strongest clues surrounding the homonym's meaning are in the story **ending**.
- The sense, context and homonym are jointly encoded with the ending.
- A **regression head** is added to predict the final **plausibility score**.

```
[CLS] Sense: ... Homonym: ... [SEP] <ending> [SEP]
```

### Direct Prompting

- We ask LLMs to solve Task 5, varying both the specific model and the prompt.
  - `Official:` The prompt provided by the organizers.
  - `Base:` We create a truncated version of the official prompt.
  - `Percent:` We request a score between 1 and 100, rescaled later.
  - `1Shot:` We prepend a fully labeled example to the instance.
- The sense, context and homonym are jointly encoded with the ending.
- A **regression head** is added to predict the final **plausibility score**.

### WSD

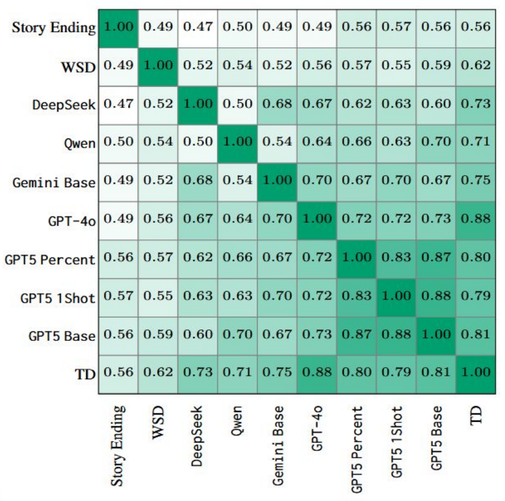
- We employ a WSD tool (ConSec), inspired by WSD's similarity to Task 5.
- Probability scores of each sense are linearly rescaled to the range [1,5].

## Analysis

Signal **variety** appears important to the ensemble:

- Non-LLM systems (Story Ending, WSD) perform worst, but are highly weighted.
- Gemini, DeepSeek, are weighted higher than their performance, suggesting varied model families improve robustness.
- Correlation confirms that LLM responses are more similar to each other than to non-LLM systems, that responses within LLM families are correlated.

**Output Correlations**




## One Homonym Per Translation

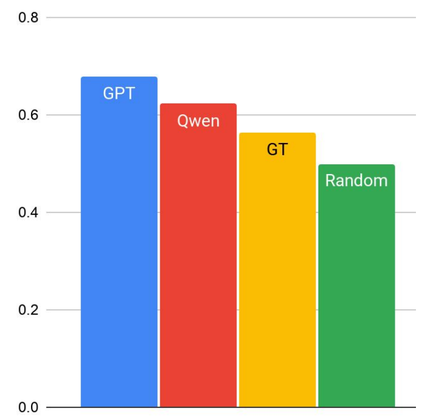
- The one homonym per translation (OHPT) principle states a homonym's meaning can be deduced from its translation.
- To use OHPT for task 5, we translate the sentence in context to determine the sense of the homonym.

She asked her friend to get rid of the **bugs**?

↓ Translate

Le pidió a su amigo que se deshiciera de los **insectos**.

**Correlation w/ Human**




- We evaluate the correlation (ROC-AUC score) between OHPT outputs and human judgement.
- We use GPT, Qwen, and Google Translate (GT) as translators.
- Correlation is consistently above random, but varies based on **translation quality**.

## Conclusion

We have demonstrated an effective approach to SemEval 2026 Task 5:

- Established that TD improves performance on Task 5 over direct prompting.
- Shown that ensembling TD with the outputs of other systems improves performance further.
- Demonstrated that classical methods such as WSD and embedding-based methods provide valuable varied signals.
- Achieved second place on the official leaderboard for Task 5.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (Amii).

UNIVERSITY OF ALBERTA

@SemEval 2026