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

Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada

amii

Task 5: The goal is to predict the **plausibility** (on a 1-5 scale) that a word expresses a given meaning in **varied contexts**, for example:

Context: Anna's room was a mess, and her computer kept crashing.

Sentence: She asked her friend to help her get rid of the **bugs**.

Varied Endings:	They were crawling on the keyboard.	(none)	New antivirus software didn't do the trick.
Sense Plausibility:			
			
			

Main Idea

- Building on two observations from prior work:
  a. LLMs struggle with numerical reasoning tasks
  b. Breaking tasks into pieces for LLMs improves robustness
- This motivates our **Task Decomposition (TD)** system
- TD breaks Task 5 into simpler tasks for an LLM to solve
- The numerical final answer is handled by a regression tool

Task Decomposition (TD)

- We prompt an LLM with 4 *binary* questions approximating Task 5
- We also prompt it directly to provide a plausibility score from 1 to 5
- We train an XGBoost regressor to produce a final plausibility score

Prompts:	Model Responses
Which of the two senses is more likely?	(**sense a**, sense b)
Is it **clear** or **ambiguous** which sense is being expressed?	(**clear**, ambiguous)
Is the given sense **definitively** correct?	(yes, **no**)
Is **at least one** meaning plausible?	(**yes**, no)
Rate how plausible the given meaning is	(1, 2, 3, **4**, 5)

Output

4.25

Additional Systems

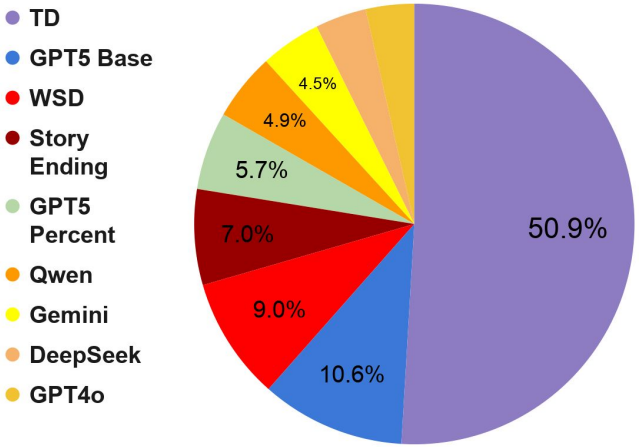
Story Ending:

- We fine-tune an embedding model to predict a plausibility score
- The sense, context and homonym are jointly encoded with the **ending**.

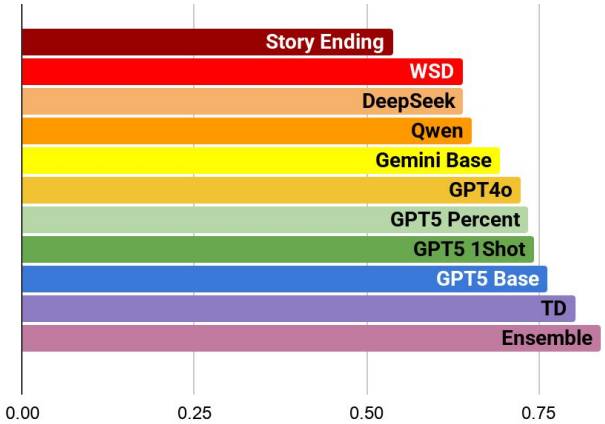
Word Sense Disambiguation (WSD):

- WSD is the task of selecting which sense a word expresses.
- We apply a WSD tool (ConSec) to Task 5, due to their similarity.
- Probabilities for each sense are rescaled to the range [1,5].

Ensemble Weights



Performance (Spearman)



Ensemble & Results

- We also ensemble TD with other systems, including direct LLM prompting and WSD.
- Our ensemble outperforms all component systems.
- Our ensemble is dominated by **TD**, our best-performing system.
- Our Non-LLM systems (WSD and Story Ending) are weighted higher despite lower performance.
- This suggests that the variety in their signals is valuable.

Contributions

An effective approach to SemEval 2026 Task 5:

- Second place on the official leaderboard
- Task Decomposition improves performance over direct prompting
- Ensembling TD with other models further improves performance
- Including non-LLM methods such as WSD boosts performance