Lingo_Research_Group at SemEval-2026 Task 9:

Evaluating Prompt Variants for Polarization Detection

Pritam Kadasi[1*], **Anuj Tiwari**[2†], **Mayank Singh**[1]

[1]Lingo Research Group, Indian Institute of Technology Gandhinagar [2]Noida Institute of Engineering and Technology, ML Collective

aj11anuj123@gmail.com {pritam.k, singh.mayank}@iitgn.ac.in

[*]Equal contribution, Corresponding author. [†]Equal contribution.

Motivation

Online polarization is a growing challenge in socio-political discourse, characterized by hostile in-group vs. out-group framing, and rhetorical dismissal. Automatic detection is difficult because polarization is frequently communicated by:

- **Sarcasm, abbreviating the ideology, framing it ironically**
- **Region and language specific cultural norms**
- **Multilingual scale across diverse languages**

Task specific fine-tuning is costly and does not scale to low-resource languages. There's prompt based inference, which's a scalable alternative, but we sought to answer the question of how much does prompt design matter?

Task Overview

SemEval-2026 Task 9 defines three subtasks of increasing granularity:

Subtask	Description
Subtask1	*Binary* - Is the text polarized or not?
Subtask2	*Polarization categories multi-label:* political, racial/ethnic, religious, gender/sexual, other
Subtask3	*Manifestation categories multi-label:* stereotype, vilification, dehumanization, extreme language, lack of empathy, invalidation

Table 1: Three subtasks of SemEval-2026 Task 9

Data spans **22 languages** including Amharic, Arabic, Bengali, Chinese, Hindi, Nepali, Urdu, English, Spanish, Turkish, and others, sourced from social media posts around elections, clashes, and sociopolitical events.

Methodology

We consider all subtasks as **instruction-following problems** and the design of the prompts as the main experimental variable (no parameter changes, no external data). Twelve variants of the prompt were developed for it starting with granular level base prompt with increase in context, defination, example and understanding:

Prompts	Design Dimension
1-2	Minimal / no task context
3-4	Short task definition (+ translate then classify)
5-6	Decision boundary clarification; conservative rule
7-8	Step-by-step reasoning before prediction
9-12	In-context labeled examples with reasoning

Table 2: Five design levels across 12 prompt variants

Aya-101 and Gemma3-27B were tested, and Gemma3-27B demonstrated a strong performance across all subtasks and prompt variants, yielding the highest F1 scores in each macro score. The prompts were all in **English** and were applied identically to all 22 languages, using the models multi-lingual features. No language specific customization was applied.

Results

Overall Performance Summary

Subtask	Lingo Macro F1	Baseline F1	Avg Accuracy	Lang.
Subtask1: Binary Detection	**0.762**	0.760	0.819	22
Subtask2: Type Classification	**0.587**	0.487	0.678	22
Subtask3: Manifestation ID	**0.444**	0.397	0.498	18

Table 3: Official test set results (Gemma3-27B, best prompt) vs POLAR baseline averaged across all languages

Lingo Research Group vs POLAR Baseline

Subtask 1				Subtask 2				Subtask 3			
Lang.	**Lingo**	**Base**	**Rk.**	**Lang.**	**Lingo**	**Base**	**Rk.**	**Lang.**	**Lingo**	**Base**	**Rk.**
Amharic	0.793	0.715	3rd	Amharic	0.546	0.372	–	Amharic	0.519	0.443	–
Arabic	0.769	0.796	–	Arabic	0.652	0.486	–	Arabic	0.533	0.390	–
Bengali	0.793	0.853	–	Bengali	0.422	0.289	**1st**	Bengali	0.161	0.087	–
German	0.732	0.671	–	German	0.599	0.408	3rd	German	0.449	0.349	–
English	0.810	0.780	–	English	0.503	0.333	–	English	0.469	0.410	–
Persian	0.719	0.842	3rd	Hindi	0.770	0.791	–	Hindi	0.719	0.235	–
Hindi	0.821	0.738	3rd	Khmer	0.694	0.627	–	Nepali	0.669	0.131	**3rd**
Italian	0.346	0.677	–	Nepali	0.805	0.722	**2nd**	Spanish	0.471	0.509	–
Nepali	0.918	0.880	**2nd**	Polish	0.625	0.449	3rd	Turkish	0.466	0.769	–
Punjabi	0.795	0.790	3rd	Spanish	0.664	0.594	–	Urdu	0.561	0.532	–
Polish	0.843	0.724	**1st**	Turkish	0.624	0.471	3rd	Chinese	0.654	0.000	–
Spanish	0.788	0.727	–	Urdu	0.798	0.713	**1st**				
Turkish	0.798	0.696	–	Chinese	0.825	0.670	3rd				
Urdu	0.816	0.789	3rd								
Chinese	0.921	0.869	–								

Table 4: Official test set results (Gemma3-27B, best prompt) vs POLAR baseline averaged across all languages and Subtasks

Official Rankings of SemEval-2026 Task 9

Team	Subtask	Total	1st	2nd
UTokyo Tsuruoka Lab	Subtask1	12	8	4
NYCU-NLP	Subtask1	12	3	5
PSK	Subtask1	9	2	4
Lingo Research Group	**Subtask1**	**5**	**1**	**1**
UTokyo Tsuruoka Lab	Subtask2	13	7	5
NYCU-NLP	Subtask2	15	6	5
Lingo Research Group	**Subtask2**	**7**	**2**	**1**
SMASH	Subtask3	16	9	4
NYCU-NLP	Subtask3	11	7	3
Lingo Research Group	**Subtask3**	**1**	**0**	**0**

Table 5: Official SemEval team rankings across all subtasks (selected teams shown). Lingo achieves 3 gold, 2 silver, 10 bronze medals total.

Analysis

Cross Subtask Performance Degradation

The F1 drop from Subtask1 → Subtask3 is due to:

1. **Multi label complexity** i.e. macro F1 errors compound with label dimensionality.
2. **Label abstraction** for example in Subtask3 manifestations like *lack of empathy* and *invalidation* lack deep level understanding.

Conservative Prompting Trade-off

The best prompt predicts polarization only when evidence is explicit. This **helps Subtask1 and Subtask2** (fewer false positives) but **hurts Subtask3** suppressing recall for subtle targets and producing misleadingly high accuracy but low macro F1.

Language Specific Observations

- **Nepali & Chinese:** use explicit group markers topical cues to excel in Subtask1/Subtask2.
- **Hindi:** directly maps vilification, extreme language to Subtask3 labels into leads.
- **English underperforms:** because of the use of irony, sarcasm, and ideological shorthand (e.g., 'deep state') fool conservative prompts.
- **Hausa & Italian:** show high accuracy but near zero macro F1 which shows systematic recall suppression.

Conclusions

Coarse grained polarization detection achieved with prompt based multi-lingual LLM models is effective as it does not require fine-tuning. It still remains a challenge to fine grained sociolinguistic classification (Subtask2/Subtask3) with the increase of label abstraction and multi label complexity. Prompt design also matters: example augmented, reasoning guided prompts with clear decision boundaries outperform minimal instruction baselines.