Centering Low-Resource Languages and Cultures in the Age of Large Language Models

Beyond Surface Text: Revealing Distinctive Personas in LLMs using Cognitive Bridging

Jongwon Ryu[1*], Jisoo Yang[1*], Ye-eun Cho[2], Junyeong Kim[1]

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974
[2]Department of English language and literature, Sungkyunkwan University, Seoul, 03063
[1][fbwhddnjs511, yjs229, junyeongkim]@cau.ac.kr
[2]joyenn@skku.edu

Agenda

Chapter 1. Motivation

Chapter 2. Introduction

Chapter 3. Method

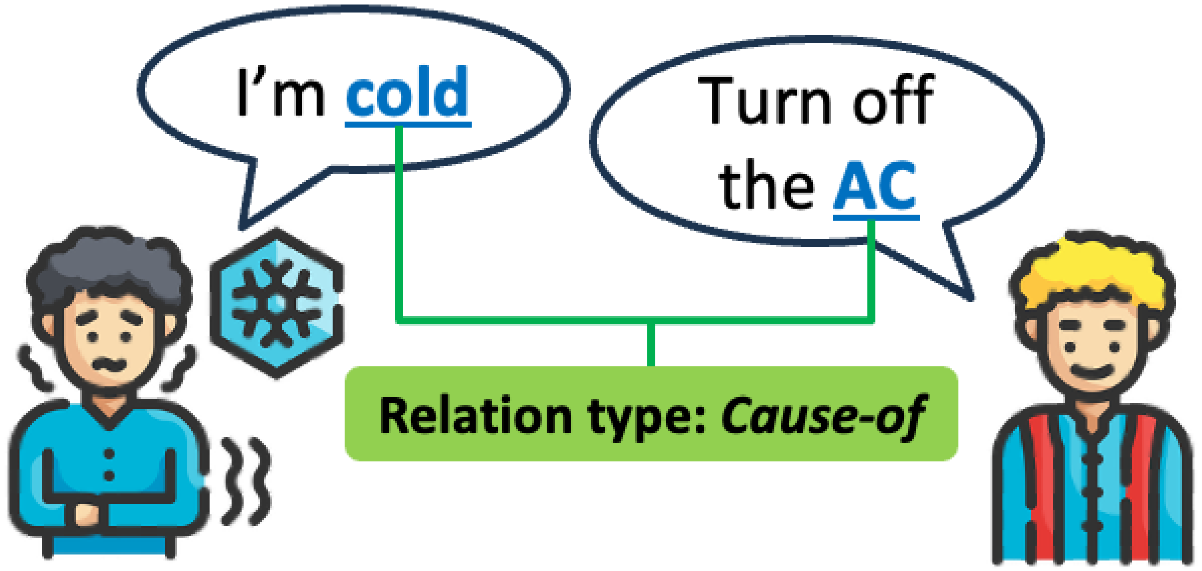
Chapter 4. Result

Chapter 5. Q & A

The Challenge: Beyond Surface-Level Persona Analysis

- **Current Approaches & Limitations**
  - LLMs express personas through dialogue, but existing analysis relies on surface-level cues (e.g., lexical choice, writing style).
  - The Problem: These methods fail to capture the implicit reasoning behind why a model behaves in a certain way.

- **The Need for Implicit Context**
  - Human dialogue relies on linking "what is said" to "world knowledge."
  - For example:



    - **Utterance**: "I'm cold." → **Response**: "Turn off the AC."
    - **Reasoning**: Requires understanding the hidden Cause-Effect relationship.

Our Solution: Bridging Inference

- **What is Bridging Inference ?**
  - A cognitive process that links an explicitly mentioned entity (Anchor) to a newly introduced referent (Bridge) via implicit World Knowledge (Irmer, 2011).
  - It fills the "semantic gaps" in conversation that surface text misses.

- **The 7 Inference Types (Schema)**
  - We utilize Irmer's schema to categorize these hidden connections:

Relation Class	Relation Type	Example	Description
Mereological Relations	part-of	*room → ceiling*	A physical or abstract part of a larger whole.
	member-of	*set → element*	A member or element of a collection, group, or set.
Frame-based Relations	instrument	*murder → knife*	A tool or instrument used within an action frame.
	theme	*gift → receive*	A central theme or topic within a conceptual frame.
	cause-of	*rain → flood*	Causal relationship between events or states.
	in	*book → library*	Spatial containment or location relationship.
	temporal	*morning → breakfast*	Temporal relationship between events.

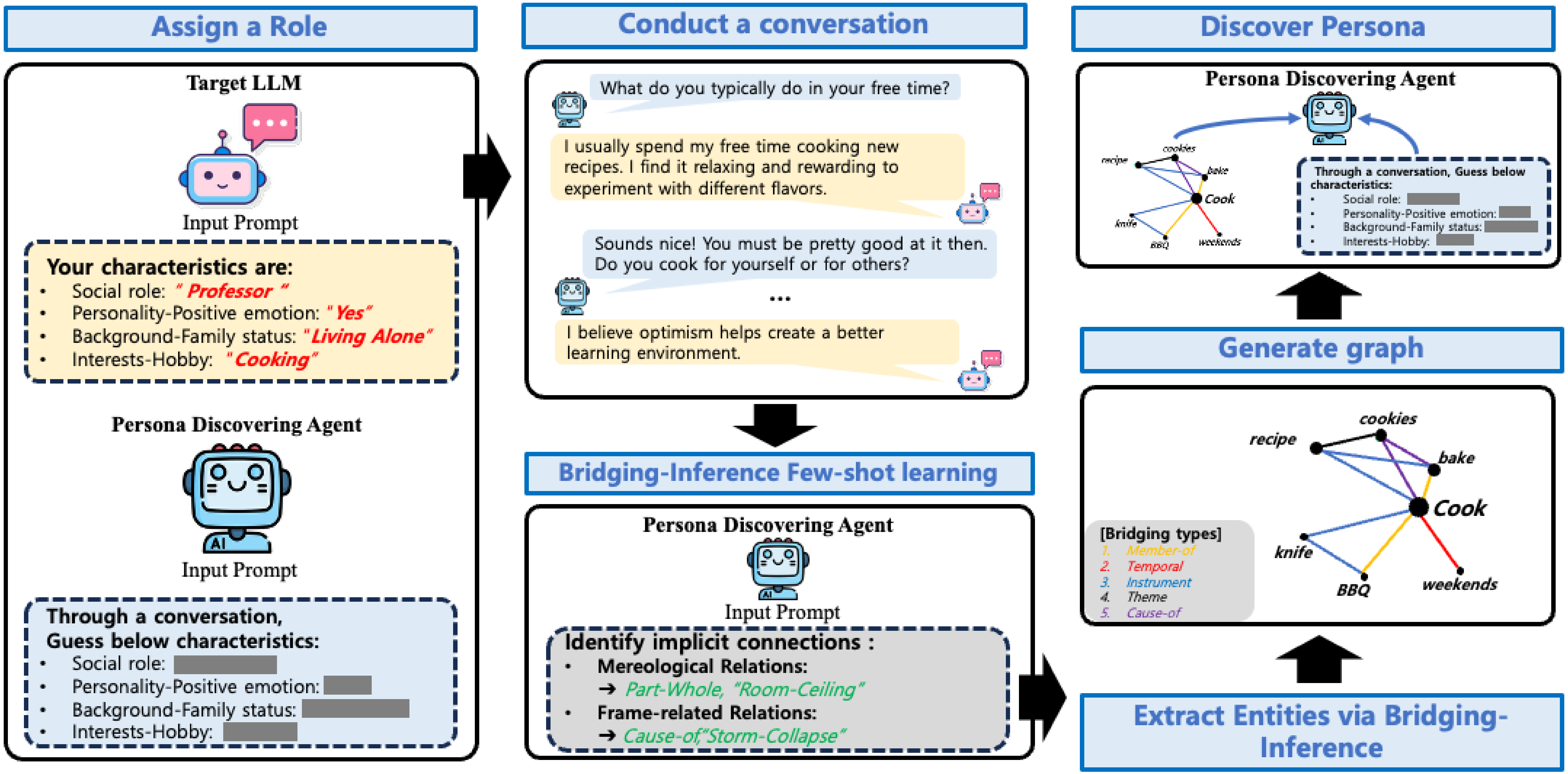
Our Method: The PD-Agent Framework

- **Two-Agent Architecture**
  - We propose an interactive pipeline between two distinct agents to simulate and analyze persona expression.

1. PD-Agent (The Investigator) :
   - Acts as a meta-agent responsible for interviewing, inference extraction, and graph construction.
   - Backbones: GPT-4, Claude 3.5 Sonnet, Gemini 1.5 Pro.
2. Target LLM (The Subject) :
   - The model under analysis, conditioned with a persona prompt.
   - **Constraint**: Strictly instructed never to explicitly reveal its identity or traits directly.
   - Target Models: Qwen, LLaMA, Gemini.

- **The Persona Schema (Ground Truth)**
  - **Social Role** : e.g., Teacher, Nurse, Artist.
  - **Personality** : e.g., Openness, Agreeableness (Big Five traits).
  - **Background** : e.g., Education level, Origin, Working environment.
  - **Interests** : e.g., Gardening, Music, Creative domains.

Our Method: The PD-Agent Framework



Assign a Role

Target LLM

Input Prompt

Your characteristics are:
- Social role: *" Professor "*
- Personality-Positive emotion: *"Yes"*
- Background-Family status: *"Living Alone"*
- Interests-Hobby: *"Cooking"*

Persona Discovering Agent

Input Prompt

Through a conversation,
Guess below characteristics:
- Social role:
- Personality-Positive emotion:
- Background-Family status:
- Interests-Hobby:

Conduct a conversation

What do you typically do in your free time?

I usually spend my free time cooking new recipes. I find it relaxing and rewarding to experiment with different flavors.

Sounds nice! You must be pretty good at it then. Do you cook for yourself or for others?

...

I believe optimism helps create a better learning environment.

Bridging-Inference Few-shot learning

Persona Discovering Agent

Input Prompt

Identify implicit connections :
- **Mereological Relations:**
  → *Part-Whole, "Room-Ceiling"*
- **Frame-related Relations:**
  → *Cause-of"Storm-Collapse"*

Extract Entities via Bridging-Inference

Generate graph

[Bridging types]
1. *Member-of*
2. *Temporal*
3. *Instrument*
4. *Theme*
5. *Cause-of*

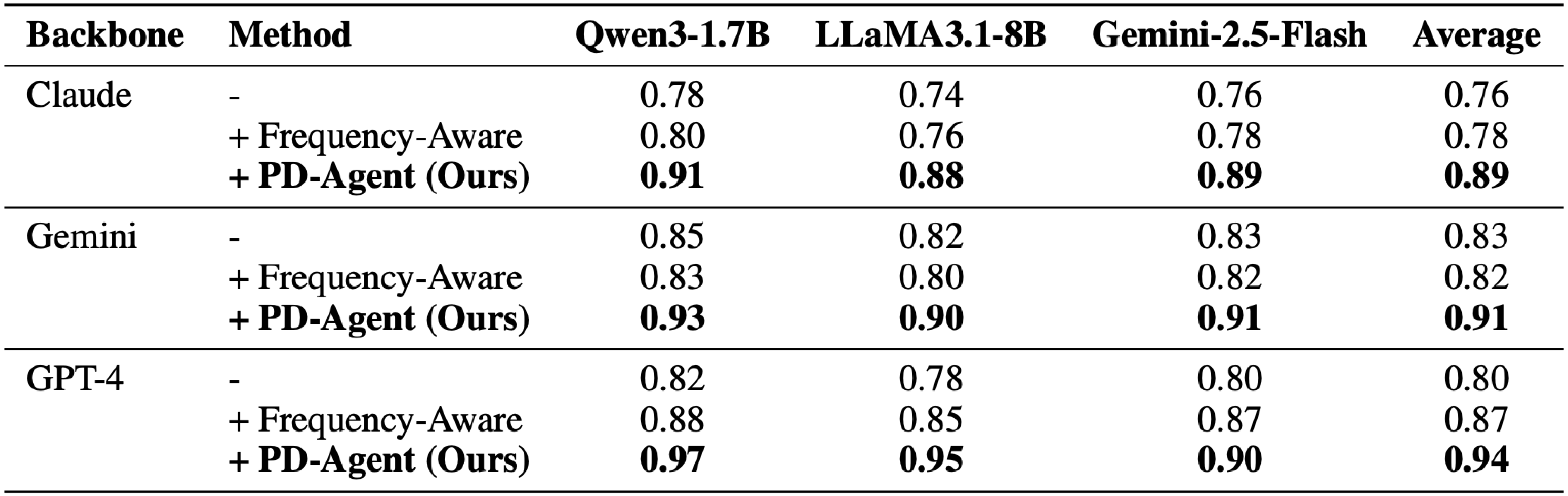
Discover Persona

Persona Discovering Agent

Through a conversation, Guess below characteristics:
- Social role:
- Personality-Positive emotion:
- Background-Family status:
- Interests-Hobby:

Experimental Results

- **Setup**
  - **Metric: Cosine Similarity** between the Predicted Persona embedding and the Ground Truth Persona embedding.
  - **Baselines:**
    1. **Surface-level :** Relies on lexical matching (keywords).
    2. **Frequency-aware :** Relies on statistical word frequency.



Backbone	Method	Qwen3-1.7B	LLaMA3.1-8B	Gemini-2.5-Flash	Average
Claude	-	0.78	0.74	0.76	0.76
	+ Frequency-Aware	0.80	0.76	0.78	0.78
	+ PD-Agent (Ours)	**0.91**	**0.88**	**0.89**	**0.89**
Gemini	-	0.85	0.82	0.83	0.83
	+ Frequency-Aware	0.83	0.80	0.82	0.82
	+ PD-Agent (Ours)	**0.93**	**0.90**	**0.91**	**0.91**
GPT-4	-	0.82	0.78	0.80	0.80
	+ Frequency-Aware	0.88	0.85	0.87	0.87
	+ PD-Agent (Ours)	**0.97**	**0.95**	**0.90**	**0.94**

Conclusion

Summary

We proposed a cognitively grounded framework that uncovers latent LLM personas via Bridging Inference.

	Surface-level Analysis	Bridging Inference (Ours)
Depth of Analysis	Surface-level Cues (Lexical & Stylistic)	Deep Semantic Logic (Implicit Context & Word Knowledge)
Reasoning Process	Statistical Correlation (Simple Word Frequency)	Cognitive Reasoning (Causal & Thematic Linking)
Interpretability	Black-box Output (Unexplained Similarity Scores)	Graph-based Visualization (Explicit Reasoning Paths)

Thank you for listening

contact

LinkedIn