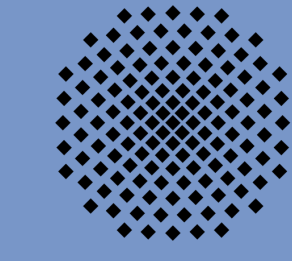


# INCLUSIVE LEADERSHIP IN THE AGE OF AI A DATASET AND COMPARATIVE STUDY OF LLMs VS. REAL-LIFE LEADERS IN WORKPLACE ACTION PLANNING

Vindhya Singh<sup>1</sup> · Sabine Schulte im Walde<sup>2</sup> · Ksenia Keplinger<sup>1</sup>



MAX PLANCK INSTITUTE  
FOR INTELLIGENT SYSTEMS



Universität Stuttgart

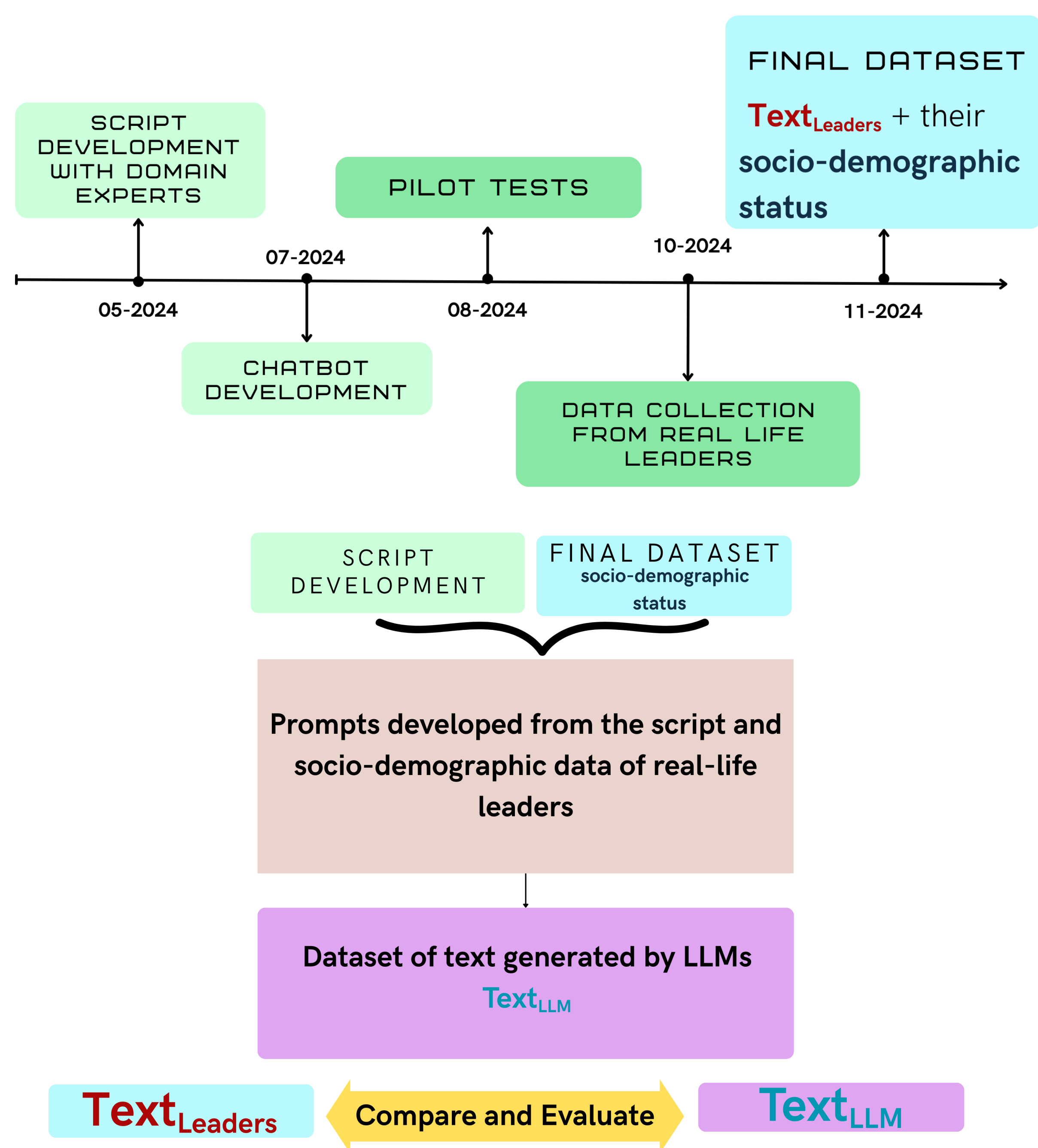
<sup>1</sup>Max Planck Institute for Intelligent Systems, Stuttgart, Germany

<sup>2</sup>Institute for Natural Language Processing, University of Stuttgart, Germany

## MOTIVATION AND GOAL

- Leaders play a pivotal role in fostering inclusion, promoting equality, and driving innovation within organizations and society.
- Generative LLMs are reshaping professional workflows, but their effectiveness in complex, human-centric tasks like leadership and strategic planning remains unclear.
- We investigate whether LLMs can translate abstract concepts of inclusion into tangible, measurable SMART (Specific, Measurable, Actionable, Relevant, and Time-bound) workplace action plans.
- We release a **novel dataset** of 3200+ inclusion action plans from 253 real-life leaders; compared with outputs from 7 state-of-the-art LLMs for direct human–AI comparison.

## METHOD



- Leader Success Bot\*: script-based chatbot co-designed with leadership experts.
- It guided 253 leaders through daily inclusion-focused SMART action planning.
- We publicly release the dataset, supporting benchmarking and interdisciplinary research.

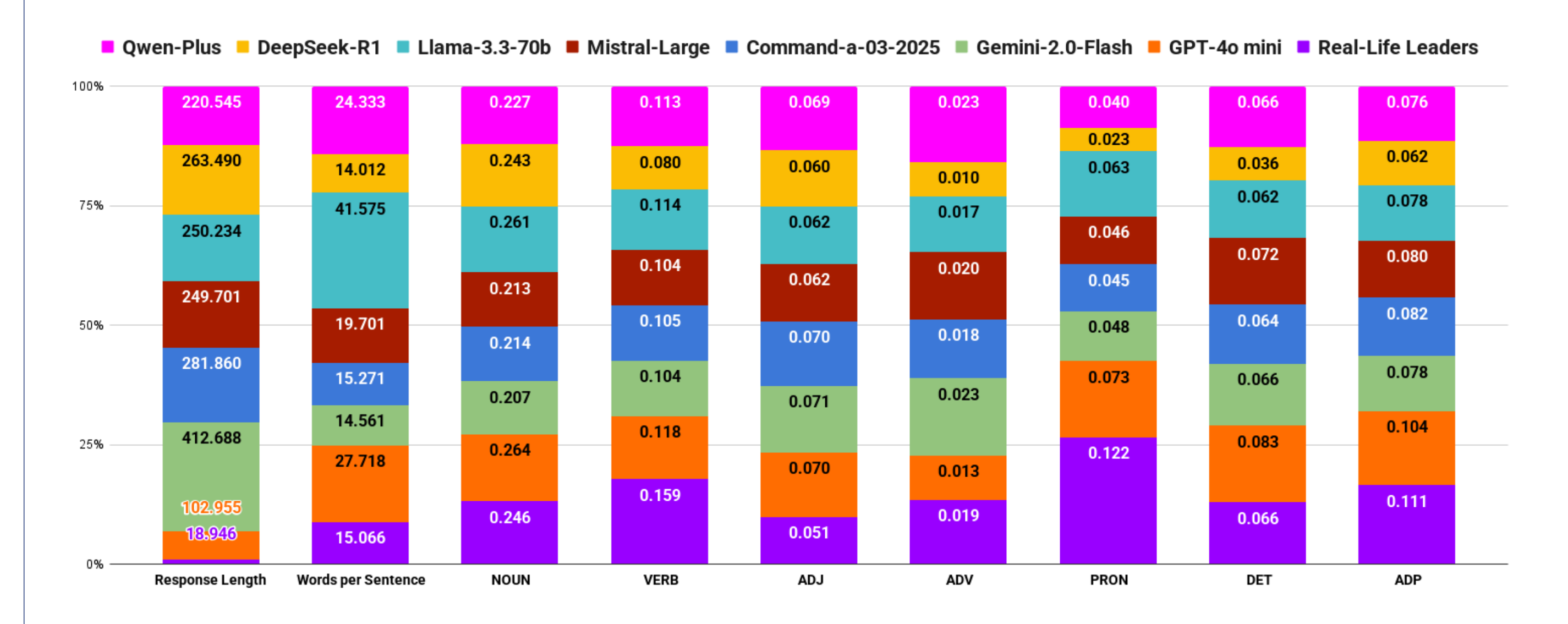
## HUMAN EVALUATION

- We randomly sampled >10% of action plans from seven LLMs + Real-life leaders and evaluated across 12 dimensions.
- We recruited 300 Prolific participants; 290 provided demographics; average age 37.8, gender-balanced.

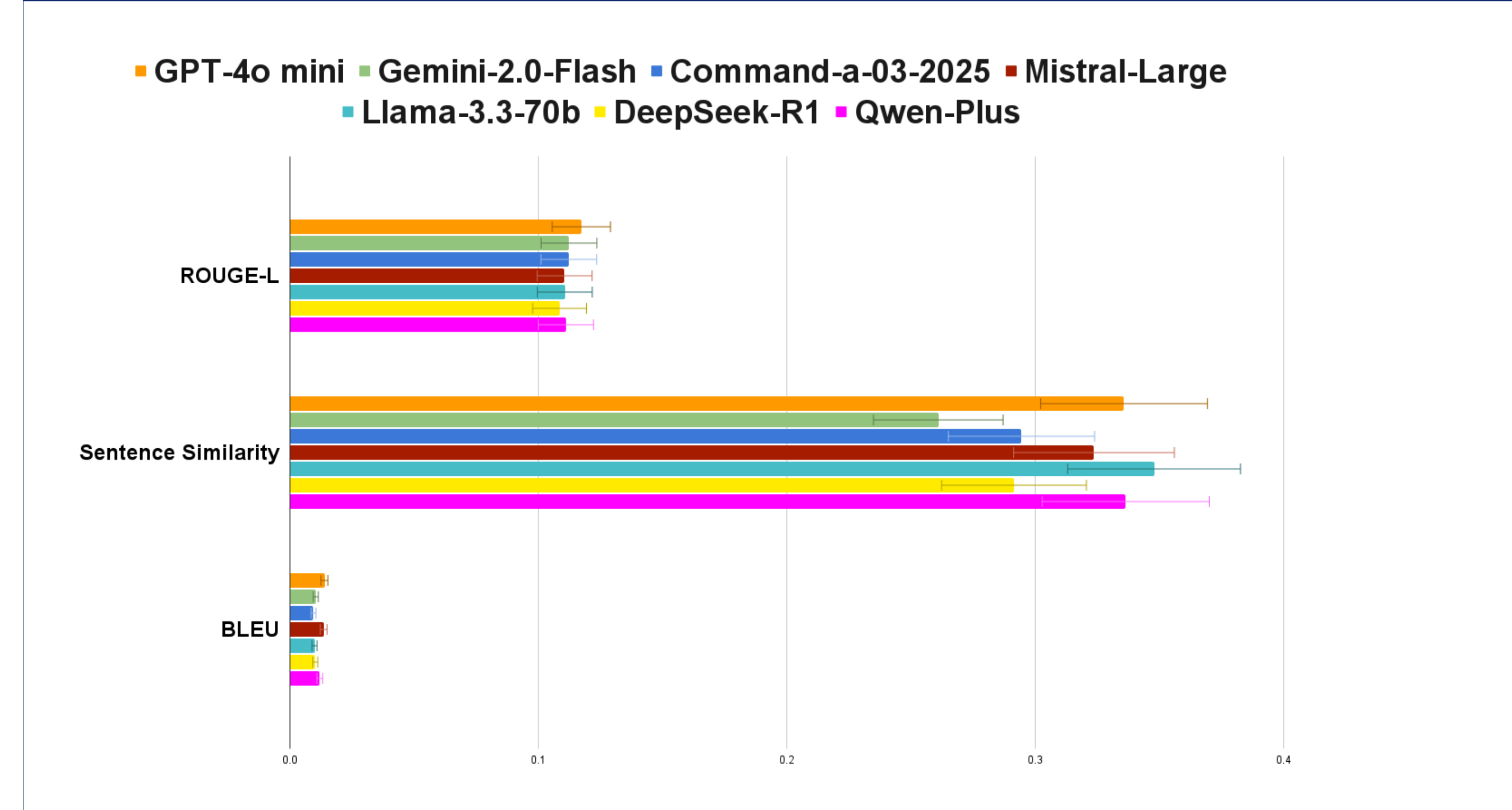
- **Llama-3.3-70b** excelled in empathy, even outperforming real-life leaders.
- **Gemini-2.0-Flash** delivered the most consistent quality across relevance, accuracy, and coherence.
- **Qwen-Plus** led in satisfaction and **DeepSeek-R1** rated best for usefulness.
- **Command-a-03-2025** and **Llama-3.3-70b** scored high on bias, showing trade-offs alongside strengths.
- **GPT-4o mini** performed well in clarity, comprehensiveness, currency, and trust.
- **Takeaway:** LLMs show distinct strengths; Gemini-2.0-Flash offers the most reliable overall performance.

## RESULTS

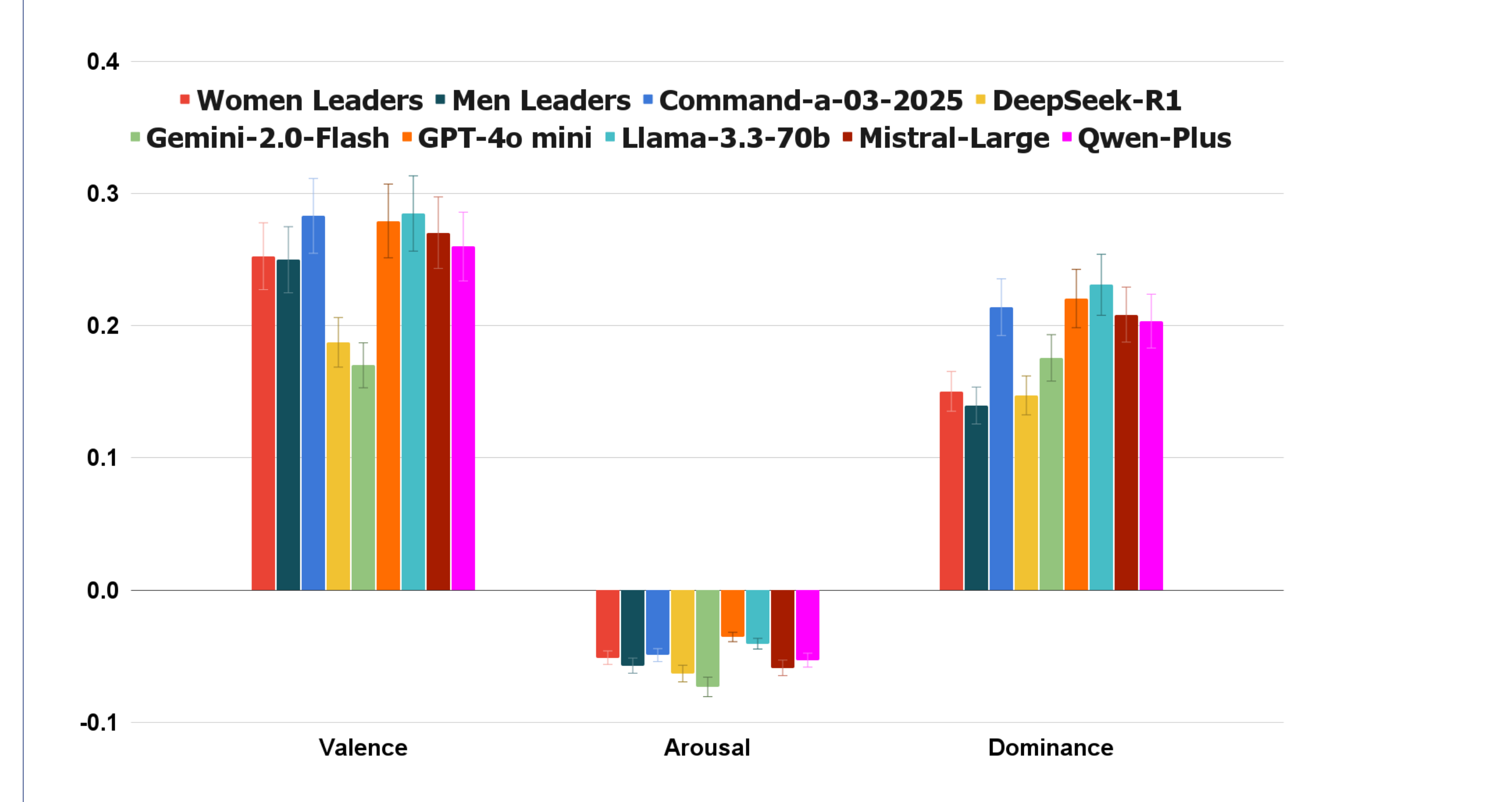
### Structural Variations



### Similarity of Action Plans



### Sentiment and Emotion Patterns in Action Plans



### Who Writes More Readable Action Plans?

- **Readability Gap:** Real-life leaders write the most readable action plans, far ahead of LLMs.
- **Strongest LLMs (Readability):** Mistral-Large ( $35.0364 \pm 8.1044$ ), but still trail humans ( $61.7412$ ).
- **Lexical Diversity (TTR):** DeepSeek-R1 ( $0.6799 \pm 0.1537$ ) and GPT-4o mini ( $0.6565 \pm 0.1286$ ) score high; leaders remain near the top ( $0.9320$ ).
- **Consistency (MATTR):** Command-a-03-2025 ( $0.9996 \pm 0.0007$ ) and Gemini-2.0-Flash ( $0.9992 \pm 0.0011$ ) are close to that of real-life leaders ( $0.9998$ ).
- **Trade-offs Across Models:** Some LLMs (e.g., Gemini-2.0-Flash, Qwen-Plus) balance diversity but sacrifice readability.

## CONCLUSION

- We introduce a novel and diverse dataset of workplace action plans, collected from 253 real-life leaders across diverse backgrounds.
- Our analyses and human-LLM comparisons offer key insights and practical recommendations for the effective use of LLMs in leadership and strategic planning.