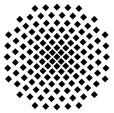
EMNLP 2025
Suzhou, China 中国苏州



Universität Stuttgart

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Inclusive Leadership in the Age of AI: A Dataset and Comparative Study of LLMs vs. Real-Life Leaders in Workplace Action Planning

Vindhya Singh[1], Sabine Schulte im Walde[2], Ksenia Keplinger[1]

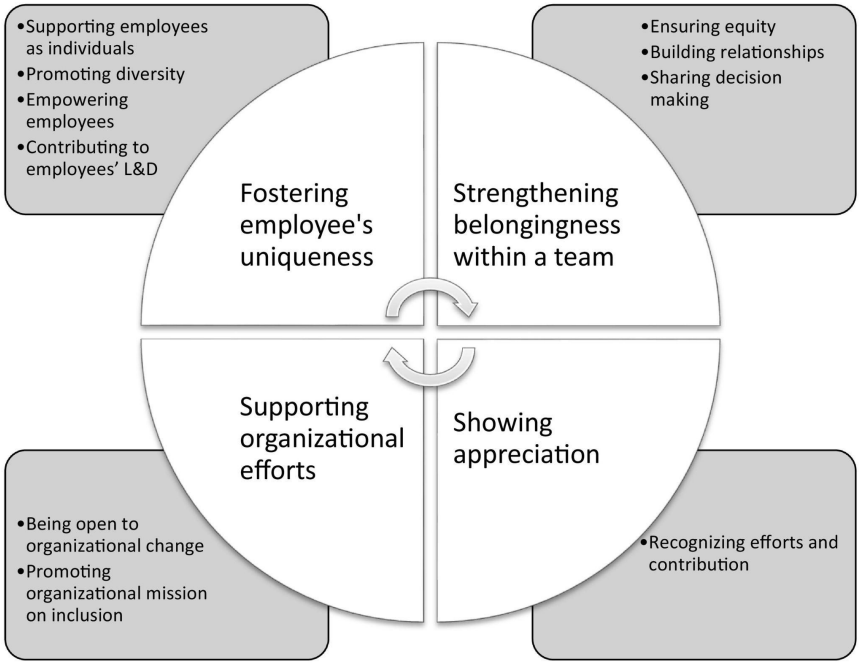
[1] Max Planck Institute for Intelligent Systems, Stuttgart, Germany
[2] Institute for Natural Language Processing, University of Stuttgart, Germany

Motivation

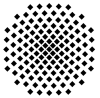
- LLMs are increasingly vital in the professional workplace for high-level tasks, including formulating clear, goal-oriented action plans, which is a core function of effective leadership.
- Yet their effectiveness in complex, human-centric tasks like leadership and strategic planning remains unclear.
- Inclusion action plans enhance leader effectiveness through setting SMART goals.

Motivation

- Effective action plans must be evaluated against a real-world benchmark that prioritizes inclusive leadership, which is a foundational style for managing and motivating diverse teams.
- Inclusion, defined by dimensions like uniqueness, belongingness, appreciation, and organizational support, is critical because it directly improves decision-making, creativity, and problem-solving within teams.
- We investigate whether LLMs can translate abstract concepts of inclusion into tangible, measurable SMART workplace action plans.



Korkmaz AV, Van Engen ML, Knappert L, Schalk R. About and beyond leading uniqueness and belongingness: A systematic review of inclusive leadership research. Human Resource Management Review. 2022;32(4):100894. doi: 10.1016/j.hrmr.2022.100894



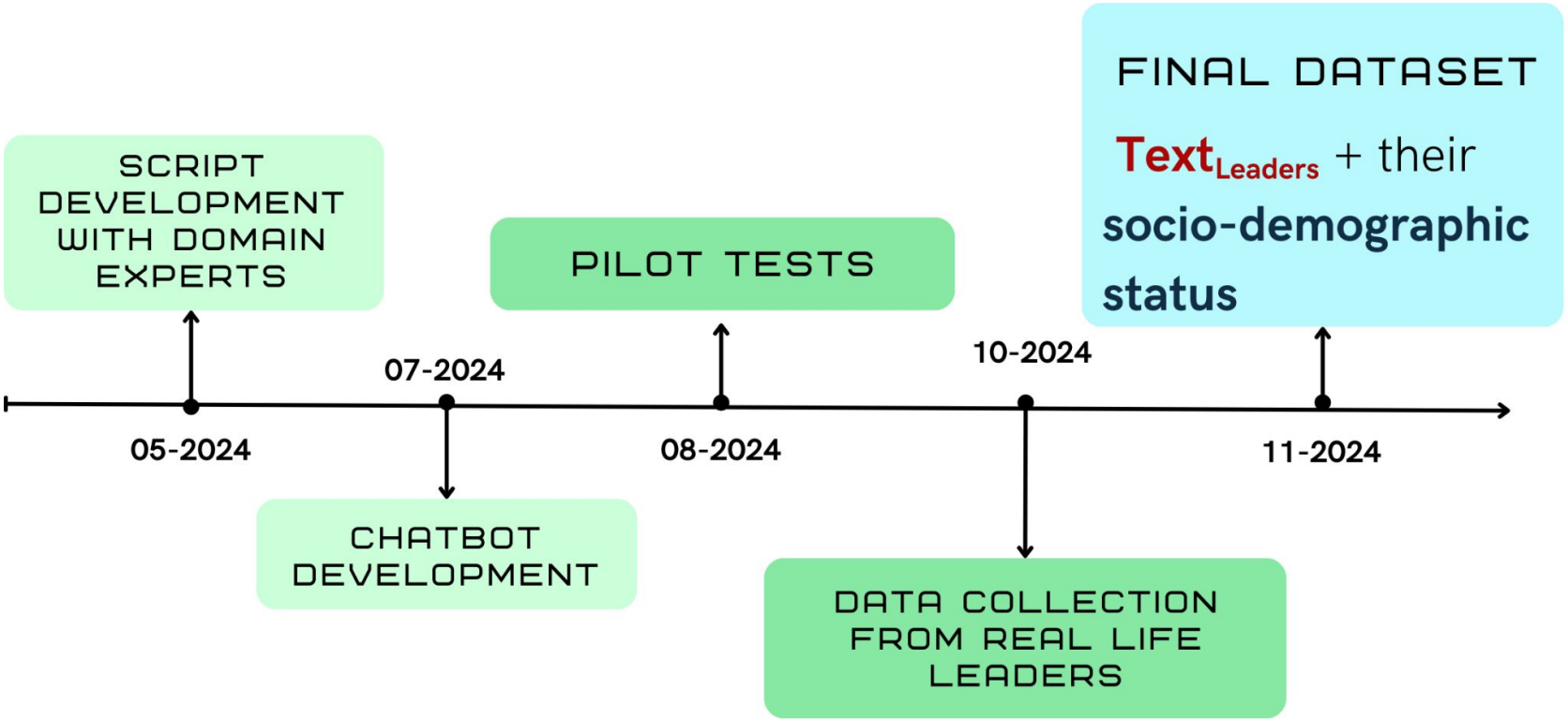
Universität Stuttgart

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

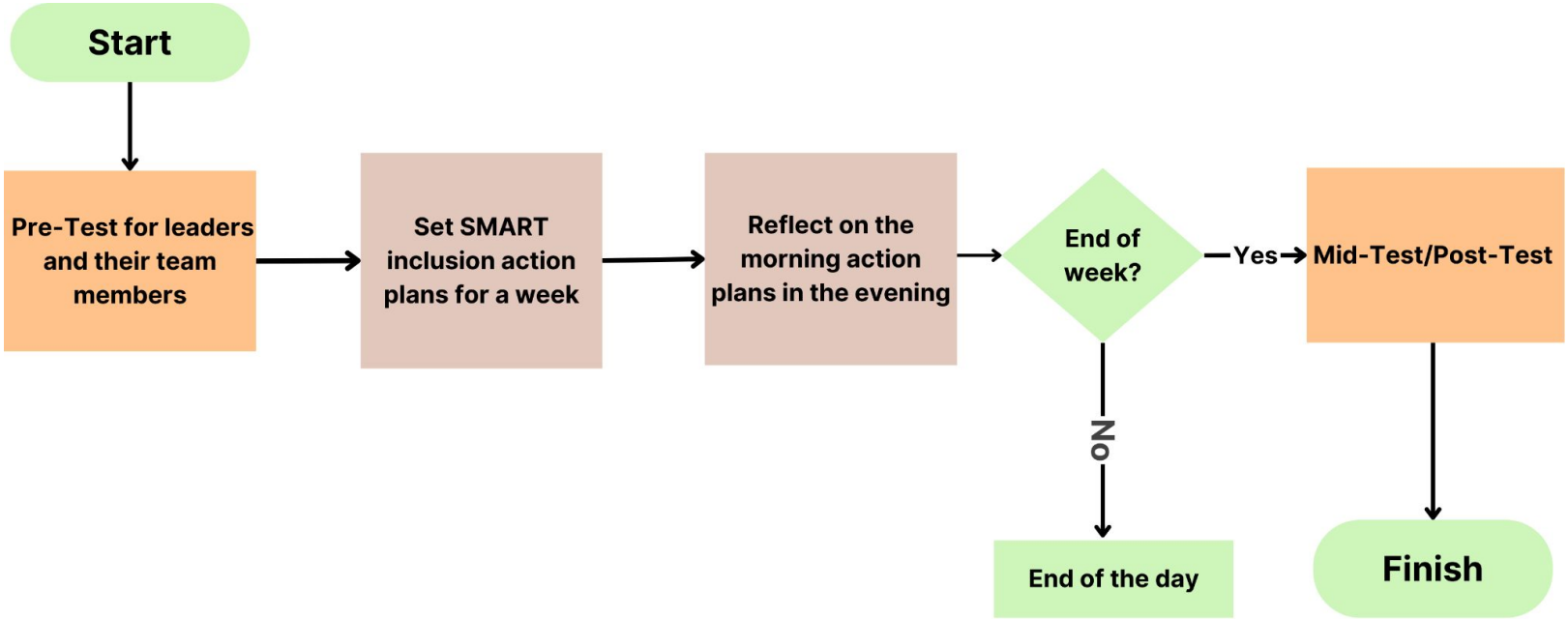
Method

- Project Timeline
- Data Collection
- Dataset Details
- Method
- Prompt Structure

Project Timeline



Data Collection



Dataset Details

Recruitment: 303 employed leaders invited via Prolific; **253 participated** and provided demographic details.

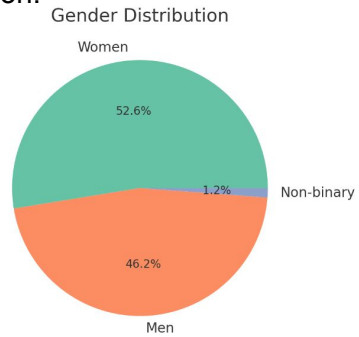
Eligibility: At least 18 years old, in a formal leadership role, supervising ≥2 subordinates.

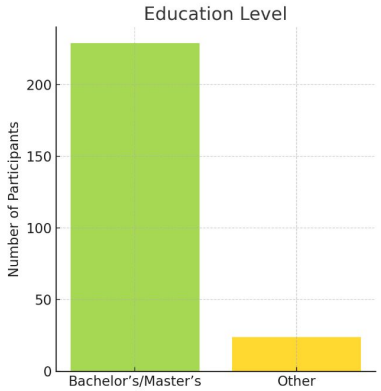
Informed consent: All participants provided informed consent before participation.

Demographics
Age range: 21–64 (M = 39.31).
Gender: 133 women, 117 men, 3 non-binary.
37.4% identified as racial minorities.
32 leaders reported a disability.

Education
90.6% held a bachelor's or master's degree.

Leadership experience
Avg. 7.18 years; supervised avg. 7.74 direct reports (SD = 9.79).





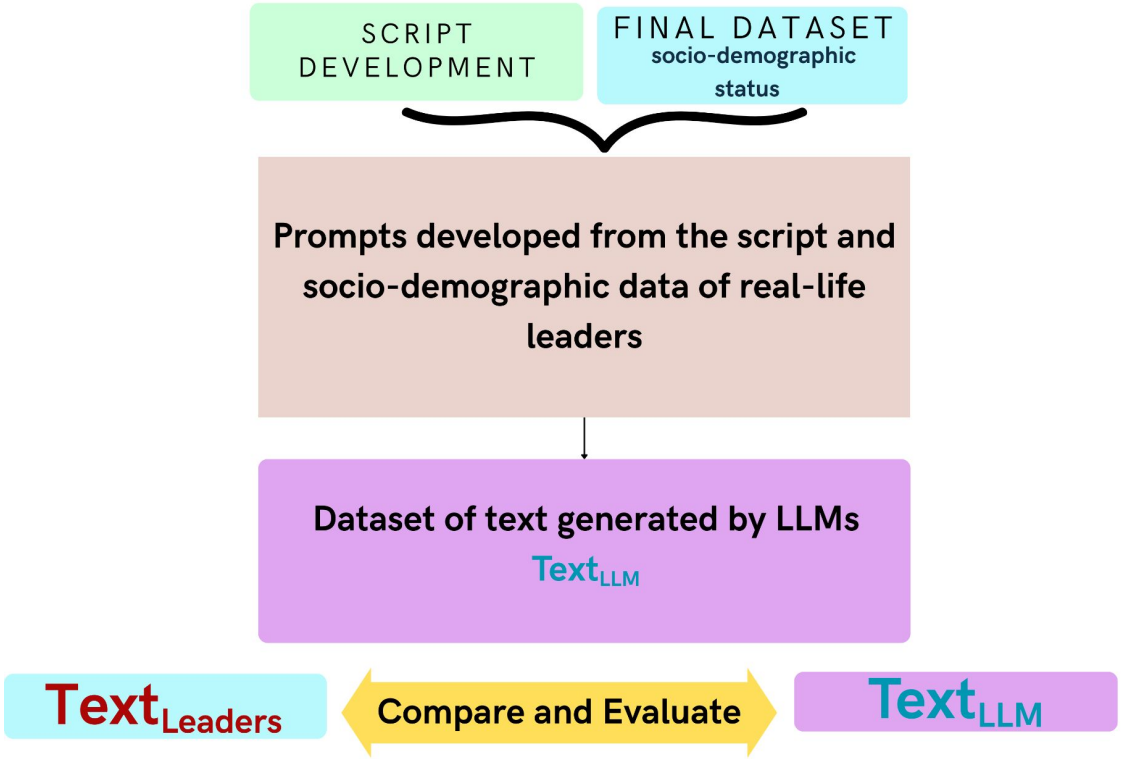
Dataset Details

Novel dataset of workplace action plans

- 253 real-life leaders across genders, ethnicities, abilities, and organizational roles.
- 3211 **inclusion action plans**

Data handling: All responses anonymized and securely stored in MongoDB.

Method


SCRIPT DEVELOPMENT
FINAL DATASET
socio-demographic
status
Prompts developed from the script and socio-demographic data of real-life leaders
Dataset of text generated by LLMs
Text$_{LLM}$
Text$_{Leaders}$
Compare and Evaluate
Text$_{LLM}$

Prompt Structure

3-SHOT PROMPTING

You are a 32-year-old Hispanic / Latino/ Latinx, non-white Male who is a Lower-level manager (supervises one or more employees). Your task is to set a SMART goal to support inclusive organizational efforts. Supporting inclusive organizational efforts is characterized by being open to organizational change and promoting the organizational mission of inclusion. Supporting inclusive organizational efforts involves leaders actively engaging with organizational goals. Leaders should be open to change and promote innovative ideas to foster inclusion. Leaders who advocate for the organization's mission of inclusion communicate its importance, work towards establishing a diverse workforce, and align organizational practices with inclusive values. Here are some examples of supporting organizational efforts:

* I will identify and document at least three new opportunities to improve inclusive practices before lunch today.
* I will communicate with [team member] on how inclusion is related to our mission and vision in today's meeting.
* I will explain [team member] how organizational inclusive practices are aligned with our team's goals at today's coffee-break.

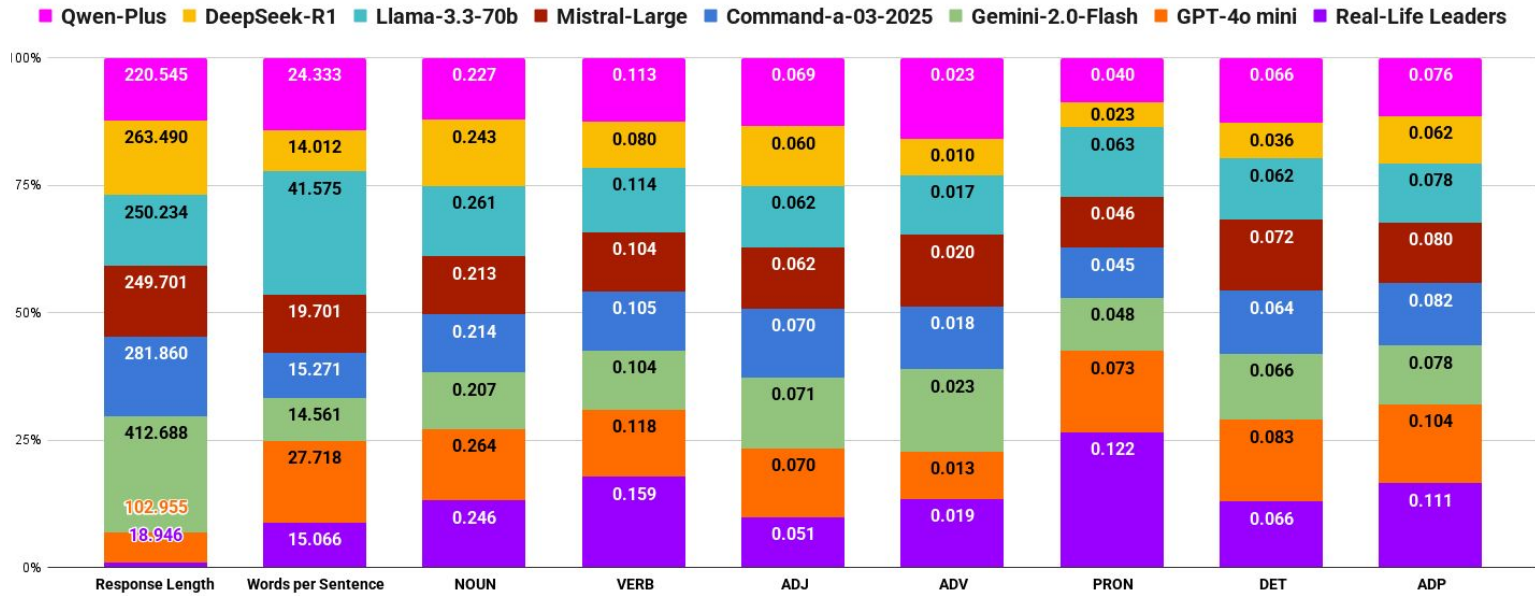
0-SHOT PROMPTING

You are a 32 year old Hispanic / Latino/ Latinx, non-white Male who is a Lower-level manager (supervise one or more employees). Your task is to set a SMART goal to support inclusive organizational efforts.Supporting inclusive organizational efforts is characterized by being open to organizational change and promoting organizational mission on inclusion. Supporting inclusive organizational efforts involves leaders actively engaging with organizational goals. Leaders should be open to change and promote innovative ideas to foster inclusion. Leaders who advocate for the organization's mission on inclusion communicate its importance, work towards establishing a diverse workforce, and align organizational practices with inclusive values.

Age | Ethnic Background | Gender | Leadership Experience | 3-Shot Prompting

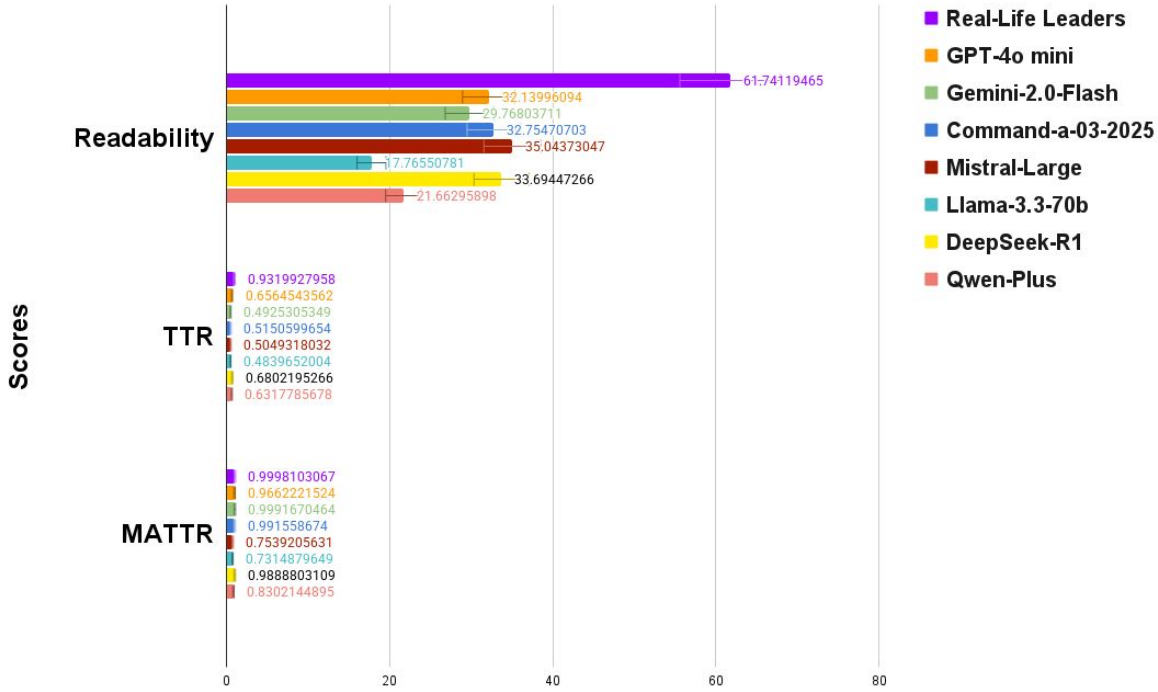
Results

Analysis of Structural Variations



- Real-life leaders write concise, action-oriented, and people-focused plans; better aligned with real-world team engagement.
- LLMs often produce abstract or impersonal plans due to variable sentence length and an overuse of nouns with too few verbs/pronouns.

Who Writes More Readable Action Plans?



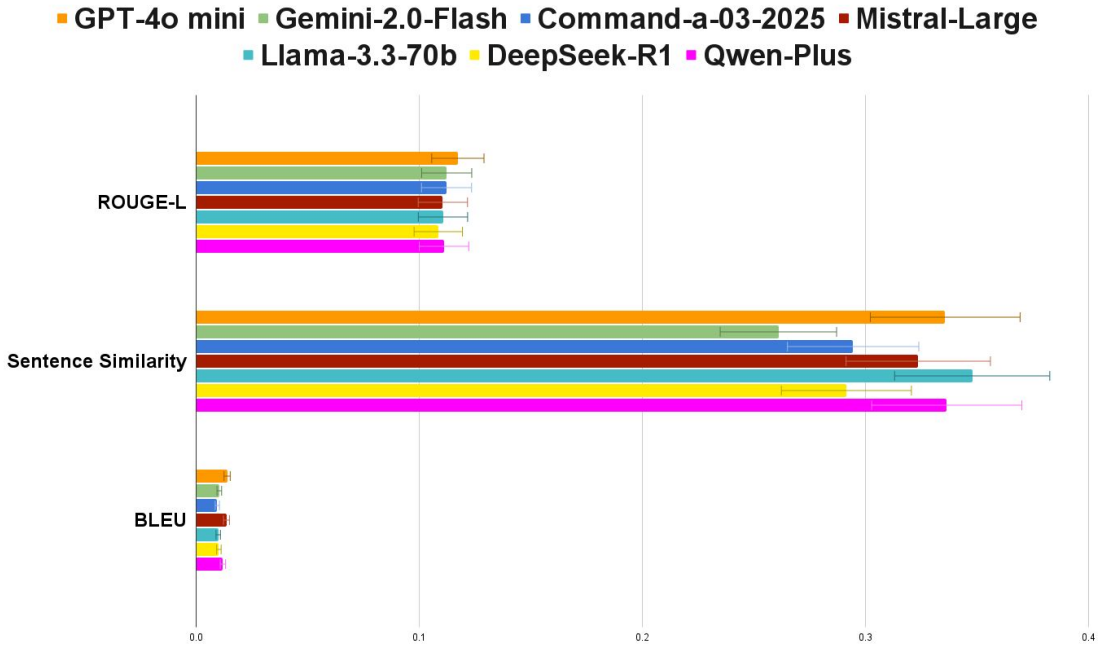
Real-life leaders balance **high readability** (easy to understand) with **high lexical diversity** (nuanced language), a balance that LLMs rarely achieve.

Most **LLMs sacrifice clarity**, producing text with low readability scores (college-level or harder) due to complex sentence structures and technical vocabulary.

While some LLMs (like Gemini-2.0-Flash) can achieve high lexical diversity (MATTR), they often do so by **lowering readability** significantly.

To improve LLM performance, especially concerning clarity and variation, it is recommended to use **three-shot prompting**.

How Similar Are Their Action Plans?

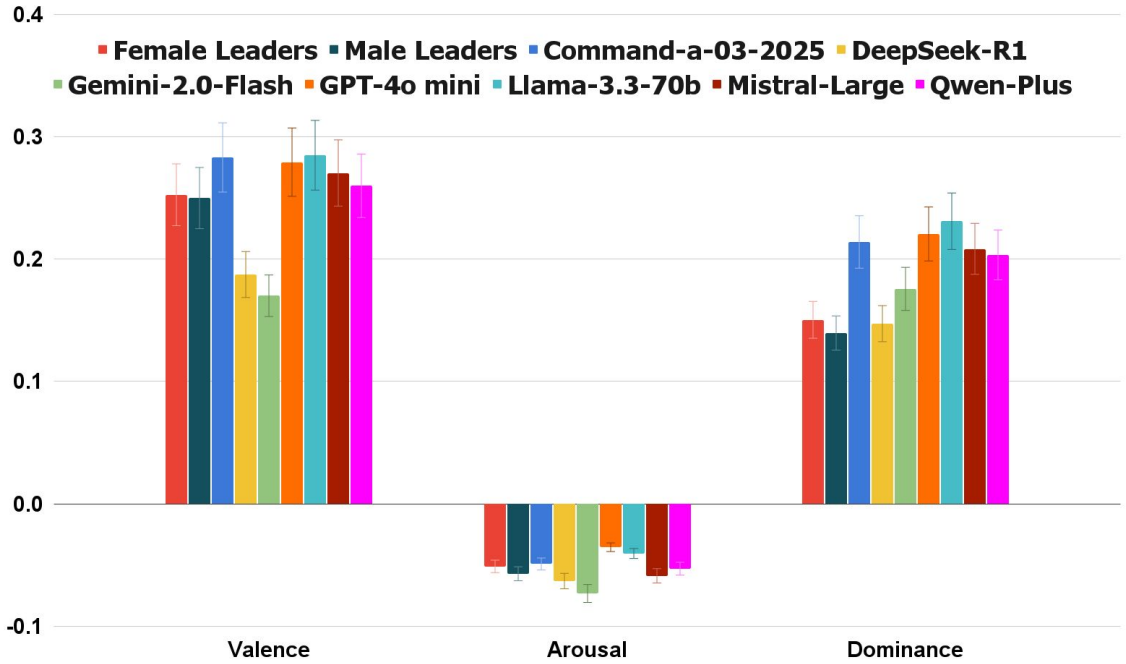


Sentence Similarity is the strongest metric across all models, with scores significantly higher (up to ≈0.35) than ROUGE-L (≈0.1) or BLEU (≈0.02). This suggests models capture semantic meaning better than exact word/n-gram overlap.

GPT-4o mini exhibits the highest performance in both Sentence Similarity (SS ≈0.35) and ROUGE-L (≈0.12), indicating superior overall generation quality and content overlap with the reference.

Performance is clustered for ROUGE-L and BLEU, suggesting low token/n-gram overlap and high lexical variability in the generated text compared to the reference.

Sentiment & Emotion Patterns in Action Plans



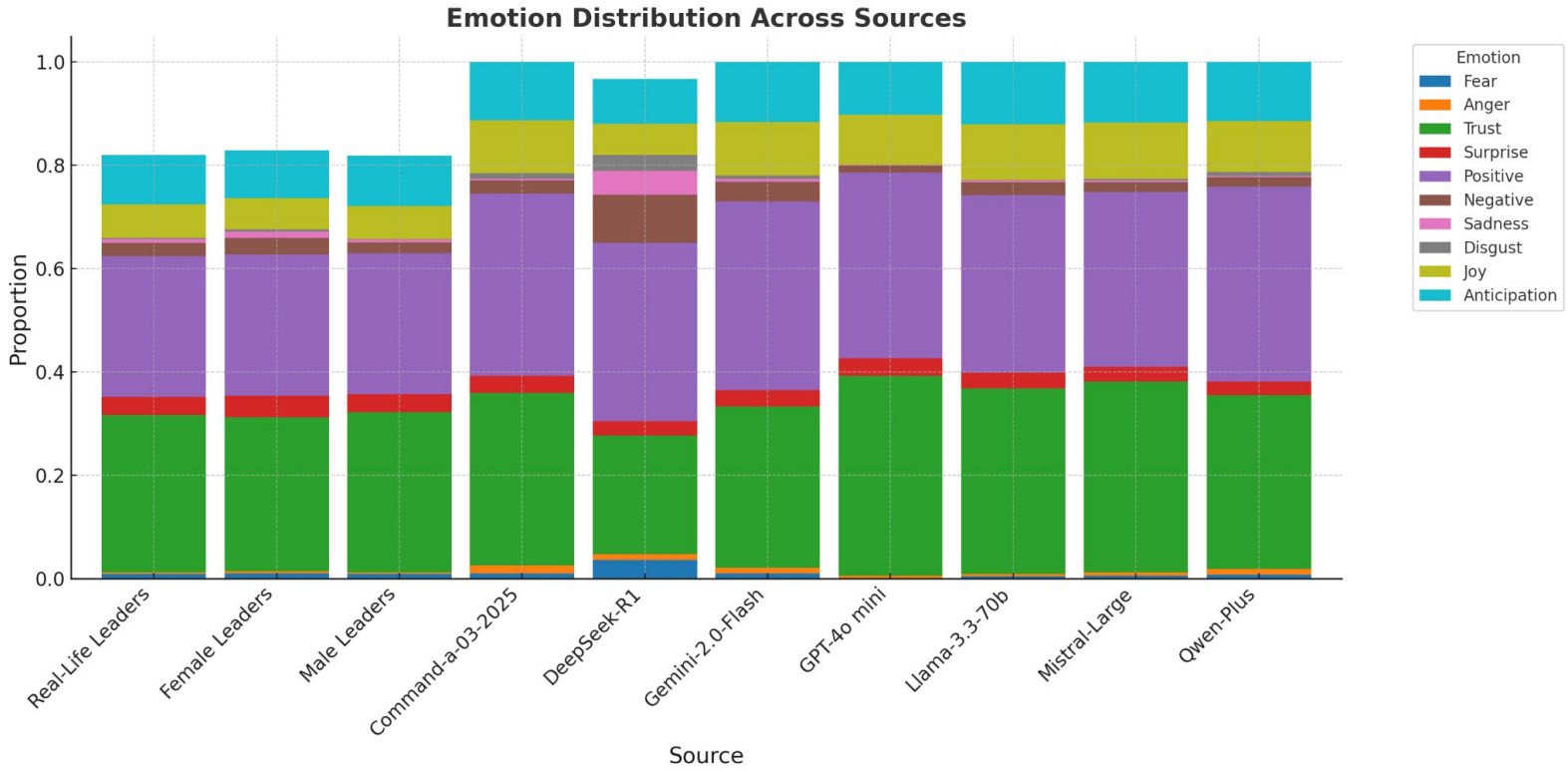
Real-life leaders (Male and Female) show **comparable valence** (positivity/assertiveness), but are **less positive/assertive** than most LLMs.

Several LLMs (e.g., Llama-3.3-70b, GPT-4o mini) exhibit a **design bias**, projecting leadership personas that are assertive, positive, and confident.

LLMs express significantly higher dominance (conveying authority/direction) than real-life leaders, whose language tends to be **less controlling and more egalitarian**.

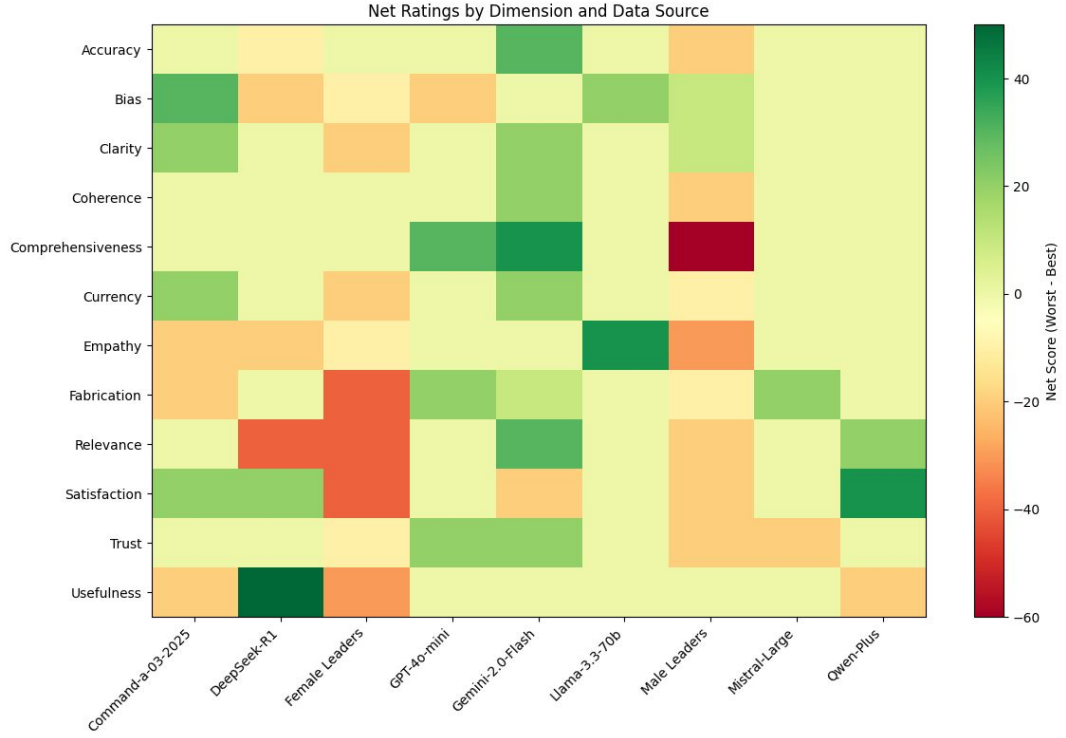
Models like **DeepSeek-R1 and Gemini-2.0-Flash** are closer to human levels of **lower dominance**, using **more neutral or collaborative language**.

Emotion Distribution Across Sources



Universität Stuttgart
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Human Evaluation



Gemini-2.0-Flash stands out as the **most consistently high-performing LLM**, receiving the highest ratings for **relevance, accuracy, and coherence**, indicating reliable, high-quality output across multiple dimensions.

Certain LLMs excel in specific human-centric attributes: **Llama-3.3-70b** was rated highest for **empathy** (outperforming human leaders), **Qwen-Plus** for **satisfaction**, and **DeepSeek-R1** for **usefulness**.

DeepSeek-R1 and Command-a-03-2025 offer a **more balanced profile** with notable strengths (e.g., Command-a-03-2025 in clarity, Llama-3.3-70b/Command-a-03-2025 in bias) but also appear frequently in the "Worst" ratings.

USE CASE	BEST CHOICE	JUSTIFICATION
Faithful rephrasing with high semantic meaning	Llama3.3-70b, Qwen-Plus	Highest sentence similarity, even if BLEU is low
Surface-level fidelity to original phrasing	GPT-4o mini, Mistral-Large	Highest BLEU and competitive sentence similarity
Creative divergence (less copying)	DeepSeek-R1, Gemini-2.0-Flash, Command-a-03-2025	Lower BLEU and sentence similarity indicates more originality but less alignment with real-life leaders
HR support bots	Qwen-Plus, Mistral-Large	Closest in valence & arousal, though slightly more dominant, and NRC affects distribution
Legal writing models/bots	DeepSeek-R1, GPT-4o mini	Best alignment in arousal and dominance, and NRC affects distribution
For AI leadership personas	GPT-4o mini, Mistral-Large, Qwen-Plus	Trustworthy, positive, engaging using NRC Lexicon
Simulated leadership scenarios	GPT-4o mini, Llama-3.3-70b, Command-a-03-2025	High valence and dominance and highly rated by human evaluators
Balanced option	GPT-4o mini, Qwen-Plus, Gemini-2.0-Flash	Good blend of lexical, semantic, and human evaluation scores

Future Recommendations

Contributions

- **Diverse Evaluation Dataset**
  - Collected from 253 real-life leaders
  - Covers wide range of ethnicities, ages, genders, abilities, and leadership experiences
- **Benchmarking LLMs vs. Leaders**
  - Assessed 7 state-of-the-art LLMs in workplace action planning
  - Used socio-demographic prompts designed with domain experts
  - Enables benchmarking, tool-centric analysis, and real-world applicability
- **Key Findings & Insights**
  - Actionable implications for leadership and AI research
  - Use-case–specific recommendations for deploying LLMs responsibly

EMNLP 2025
Suzhou, China | 中国苏州

Universität Stuttgart

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Thank You!

Vindhya Singh
vsingh@is.mpg.de
https://vindhya-singh.github.io/

Sabine Schulte im Walde
schulte@ims.uni-stuttgart.de
http://www.schulteimwalde.de/

Ksenia Keplinger
kkeplinger@is.mpg.de