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
Correspondence: vsingh@is.mpg.de, schulte@ims.uni-stuttgart.de, kkeplinger@is.mpg.de

Abstract

Generative Large Language Models have emerged as useful tools, reshaping professional workflows. However, their efficacy in inherently complex and human-centric tasks such as leadership and strategic planning remains underexplored. In this interdisciplinary study, we present a novel dataset and compare LLMs and human leaders in the context of workplace action planning, specifically focusing on translating the abstract idea of inclusion into actionable SMART goals. We developed the Leader Success Bot, a script-based chatbot co-designed with domain experts, to guide more than 250 real-life leaders in generating inclusive workplace action plans. We systematically prompted seven state-of-the-art chat-based LLMs to perform the same task using the socio-demographic data of real-life leaders and instructions co-developed with domain experts. Our publicly released dataset enables direct comparison between human and LLM-generated workplace action plans, offering insights into their respective strengths, biases, and limitations. Our findings highlight critical gaps and opportunities for LLMs in leadership applications, fostering interdisciplinary collaboration and NLP applications.

1 Introduction

Generative Large Language Models (LLMs) are now widely used in both daily life and professional settings. At the workplace, they are extensively used for a variety of tasks, including seeking advice, exploring new topics, and managing communication workflows (Brachman et al., 2024). With the rising importance of LLM usage in professional work settings, researchers have begun to investigate how LLMs can support complex, high-level functions such as human resource management, idea generation, and workplace communication (Chiarello et al., 2024). An important area where LLMs are being used in the workplace is in assisting individuals and teams in formulating clear, goal-oriented action plans, as it is a cornerstone of effective leadership. Research shows concrete, trackable, and specific goals help people stay focused, assess progress, and adapt their behavior to reach targets (Locke and Latham, 2002). However, to evaluate the effectiveness of LLM-generated action plans and compare them meaningfully with those crafted by humans, we first need a benchmark rooted in real-world leadership behavior.

Thus, we present a novel dataset comprising action plans authored by over 250 real-life leaders. We systematically compare these human-written plans with those generated by leading LLMs (GPT-4o mini, Gemini-2.0-Flash, Command-a-03-2025, Mistral-Large, Llama-3.3-70b, DeepSeek-R1, and Qwen-Plus), focusing on their ability to generate SMART (Specific, Measurable, Actionable, Relevant, and Time-bound) action plans to foster inclusion in teams. Inclusion refers to deliberate efforts that ensure individuals from diverse backgrounds feel respected, valued, and empowered to participate fully in organizational activities. Unlike diversity, which emphasizes representation, inclusion engages everyone in decision-making and team processes (Shore et al., 2011). Researchers have shown that teams make better decisions, generate more creative ideas, and solve complex problems more effectively when individuals feel safe to contribute, express their views, and participate in organizational efforts (Carmeli et al., 2010; Choi et al., 2015; Lacerenza et al., 2017; Nembhard and Edmondson, 2006; Dawson et al., 2024; Macari et al., 2024; Van Knippenberg et al., 2020). Neglecting inclusion reduces employee engagement and long-term growth (Dwertmann and Boehm, 2016). The impact of inclusion reaches beyond internal operations. It shapes how organizations influence broader social systems and public discourse (Boekhorst, 2015; Hoobler and Brass, 2006).

In workplaces, inclusion operates at both interpersonal and organizational levels (Guillaume et al.,

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19732–19753
November 4-9, 2025

2017; Nishii, 2013), and language plays a crucial part in delivering the right message across team members. As it is challenging to transform an abstract concept such as inclusion into actionable results, we collaborated with two domain experts to develop a step-wise training process to inform and educate real-life leaders on creating inclusion action plans. At the same time, researchers have called out the need to integrate technology as a 'jolt' to modify interactions and dynamics at workplaces (Wellman, 2017). Consequently, chatbots are commonly used as a tool in various sectors (Kung et al., 2023; Nasseri et al., 2023). Additionally, Gallegos et al. (2024) has highlighted the need to curate inclusive data based on community-centered frameworks.

A focus on inclusive leadership is critical as it is a foundational leadership style that directly shapes how leaders give feedback, resolve conflict, and motivate diverse teams (Randel et al., 2018; Panicker et al., 2018). Accordingly, we build our interdisciplinary project by using the **four dimensions of inclusion**, that is: (1) recognizing employees' sense of individuality (uniqueness); (2) strengthening belongingness (feeling like an esteemed member of the team); (3) showing appreciation, and (4) supporting organizational efforts, as conceptualized by Korkmaz et al. (2022). We use their proposed four-dimensional framework of inclusive leadership, from the domain of Organizational Behavior, to create a rich dataset with our domain experts and then compare it against seven state-of-the-art LLMs in the NLP domain (Mayer et al., 2025). It is important to note that though many evaluation datasets exist and have been used to study and evaluate domain adaptation of LLMs (Aycock and Bawden, 2024), to our knowledge, our inclusive dataset is novel for the leadership and organizational behavior domain.

In order to bridge the gap between theoretical LLM capabilities and applied leadership contexts, offering insights into their real-world usability rather than artificial testing scenarios, we utilize a script-based chatbot to avoid inducing AI bias in leader training and LLM-prompting parts (Raub, 2018; Vicente and Matute, 2023).

Our contributions are threefold:

- A comprehensive and diverse evaluation dataset, compiled from over 250 real-life leaders representing a wide range of ethnicities, age groups, genders, abilities, and leadership experiences.

- Assessment of seven LLMs in comparison to real-life leaders in the context of workplace action planning. This evaluation utilizes socio-demographic prompts designed with domain experts[1], enabling performance benchmarking, tool-centric analysis, and evaluation of real-world applicability.

- An in-depth discussion of key findings, actionable insights, implications, and use-case-specific recommendations for LLMs, contributing valuable perspectives to interdisciplinary research.

2 Related Work

Interactive chatbots are utilized in various contexts to facilitate external and internal communication, conduct employee training, offer customer services, generate ideas, foster well-being, help with recruitment (Fitzpatrick et al., 2017; Koivunen et al., 2022; Zhou et al., 2020) and facilitate teamwork (Avula et al., 2018; Xiao et al., 2019). While some researchers have explored language style (Elsholz et al., 2019) and identified appropriate register for chatbot design (Chaves et al., 2019), few chatbots address the topic of diversity, equity, and inclusion (for an exception see (Heo and Lee, 2019)). To ensure bias is curtailed in the design of language models, it is important to first identify it. Researchers have analyzed leadership-related gender biases perpetuated by LLMs (Newstead et al., 2023). However, much of the analysis has been limited to a few handpicked LLMs (Chisca et al., 2024; Harel-Canada et al., 2024). To address this, we examine a set of seven state-of-the-art LLMs in our study. Further, Choi et al. (2024) highlighted the trade-off between task efficiency and biased analysis when using LLMs in a domain-specific manner. Prior research on domain-adaptation of LLMs (Zhang et al., 2023; Goyal et al., 2023; Van Veen et al., 2024) has focused mostly on the domains of news, science, medicine, and government, to our knowledge. We aim to extend this work to the Organizational Behavior and the HR domain by comparing the real-world usability of LLMs and actual leaders in workplace action planning.

[1]The domain experts (Dr. Ksenia Keplinger and Dr. H. Phoenix Van Wagoner) are leading researchers in the field of Organizational Leadership and Diversity.

3 Method

3.1 Data Collection

We developed the Leader Success Bot,[2], a conversational chatbot designed to cultivate daily inclusion action plans over a two-week period. The details of the design and implementation of the chatbot from a technical perspective, including front-end and back-end infrastructure, are provided in Appendix A. We launched the Leader Success Bot in a snowball sample of 48 fully employed MBA students from a university in the Western United States over a period of 12 work days for our pilot tests and validation process. The aim of the pilot study was to ensure the usability and effectiveness of the chatbot-mediated design in real-world settings. Participants were recruited from courses in Leadership and Organizational Behavior and received extra credit for their participation. Leaders needed to: (1) interact with the chatbot for two consecutive work weeks, (2) develop SMART inclusion action plans every morning, (3) assess their plan implementation every evening, and (4) fill out three Qualtrics evaluation questionnaires (before, halfway through, and after the intervention, administered via the chatbot). A SMART **inclusion action plan** for leaders is a clear, goal-oriented procedure, outlining steps to develop leadership skills, manage teams effectively, and drive organizational success (for example, *"I will praise the technical achievements of [team member] at today's meeting."*). Participants were asked to report their weekly goal accomplishment and team inclusion climate. Forty-eight leaders provided a total of 3237 chatbot text messaging responses (of these, 432 responses were inclusion action plans) in addition to responses collected via Qualtrics surveys. These initial findings suggested that our Leader Success Bot is an effective, well-received tool for fostering workplace inclusion by successfully helping real-world leaders develop actionable inclusion plans.

Once we validated the effectiveness of the Leader Success Bot in the pilot study, we launched it for data collection from real-life leaders (Figures 1 and 2). An international, gender-diverse group of employed leaders was recruited via the Prolific platform to interact with the Leader Success Bot. Eligible participants had to be at least 18 years old, hold a formal leadership role, and supervise at least

[2]Code for chatbot design: https://github.com/Vindhya-Singh/chat-bot.git

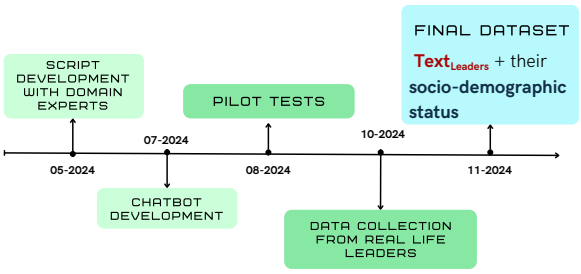

Figure 1: Timeline of data collection for this paper.

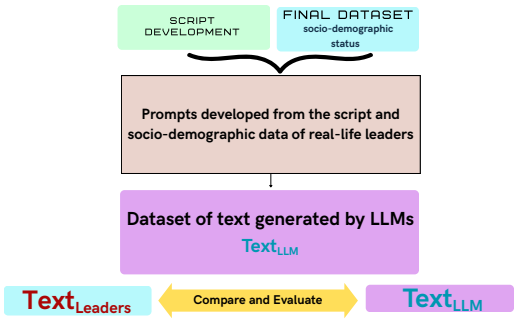

Figure 2: Methods used in this paper.

two subordinates. Out of 303 invited leaders, 253 provided demographic details and engaged with the chatbot. These leaders supervised an average of 7.74 direct reports ($SD = 9.79$) and ranged in age from 21 to 64 ($M = 39.31$). Among them, 37.39 percent identified as racial minorities, while the gender distribution included 133 women, 117 men, and 3 non-binary individuals. In addition, 32 leaders reported having a disability. The group was highly educated, with 90.6 percent holding a bachelor's or master's degree. The leaders represented various organizational levels: 39 percent were mid-level managers, 10 percent were senior or executive leaders, and roughly half were low-level managers. The participants had 7.18 years of leadership experience, on average. We anonymized all data and stored it in MongoDB. Interactions with the Leader Success Bot were saved as text, resulting in 3211 inclusion action plans from 253 real-life leaders.

3.2 Socio-Demographic Prompting

Our research aims to compare inclusion action plans created by LLMs with real-life leaders using the same instructions and examples. Domain experts helped us develop a script to train users on the abstract idea of inclusion and transform it into actionable strategies. To ensure consistency and uniformity in providing instructions to both humans and LLMs, we used prompts based on sociodemo-

graphic data from 253 real-life leaders, along with the script developed by domain experts (see structure in Table 1 and examples in the Appendix B, Figure 10). As suggested by Beck et al. (2024), we paid attention to using sociodemographic prompting to study LLM alignment by collaborating with domain experts. We refrain from prompt-tuning as: (1) our aim is to assess how current LLMs respond to identical prompts when compared to real-life leaders, creating a direct performance baseline, (2) unlike classification tasks with predefined "correct" answers, our focus is on practical utility, that is, examining how effectively these models function as assistive tools for leaders and HR professionals, (3) we evaluate LLMs as used in day-to-day tasks, mirroring how leaders actually engage with these systems, that is, conversationally and without fine-tuning, to reflect genuine user experiences (Mayer et al., 2025). For our research, we use two settings: zero-shot prompting and three-shot prompting. All seven LLMs used in our study were prompted using their respective APIs (Table 4). Each LLM was prompted 1012 times (253 leaders * 4 dimensions of inclusion), respectively, for zero-shot and three-shot prompting. Therefore, overall, we used 7084 (1012*7) action plans from seven LLMs respectively for zero-shot and three-shot prompting.

Table 1: Prompt structure for three-shot prompting.

Prompt Structure
You are a Hispanic / Latino/ Latinx, non-white 32-year-old Male Lower-level manager (supervises one or more employees). [Followed by the script with domain experts.] [Followed by 3 examples of setting SMART action plans.]

Color Legend:

- Ethnic Background
- Age
- Gender
- Leadership Level

4 Evaluation of Action Plans Generated by LLMs and Real-Life Leaders

We conduct a comprehensive linguistic analysis of action plans produced by LLMs (Cohere, 2025; DeepSeek-AI, 2025; Google, 2024; OpenAI, 2024; Meta, 2024; Mistral, 2024; Alibaba, 2025) and real-life leaders,[3] comparing structural variations, readability, sentence-level similarity, sentiment, and emotional patterns to systematically evaluate their similarities and differences across multiple dimensions of language use. For all these methods, except for human evaluation, we used the standard Python libraries and packages.[4]

4.1 Analysis of Structural Variations

Analysis: We analyze structural variations in the action plans generated by LLMs and real-life leaders (Muñoz-Ortiz et al., 2024), such as response length, words per sentence, and Part-of-Speech (POS) distribution.

Results: In Figure 3, we see real-life leaders use shorter, clearer action plans, with an average response length of 19 words, while all the LLMs generate longer action plans, with their average response length ranging from 70 to more than 400 words. Diving deep into the number of words used per sentence, we clearly see that real-life leaders prioritize brevity while LLMs tend to be more explanatory and lengthier in their generated responses. Overall, in comparison to the real-life leaders, GPT-4o mini has a well-balanced POS usage, with more nouns and adjectives, indicating instructional action plans (Mendhakar and H S, 2024). The action plans generated by GPT-4o mini have fewer verbs compared to real-life leaders, making it less action-oriented. DeepSeek-R1 has the lowest verb usage among all sources in generating action plans, suggesting less dynamic text. Gemini-2.0-Flash and Llama-3.3-70b show extremes. While Gemini-2.0-Flash is the most verbose (average response length: 412 words), Llama-3.3-70b generated the longest sentences with an average of 41 words per sentence and high noun usage, indicating likely dense, complex action plans. Notably, even though DeepSeek-R1 has a longer response length (263) than real-life leaders and most LLMs, it is the most concise in sentence length (14 words/sentence) and has a fairly balanced POS distribution, suggesting fragmentation or bullet-like structure in the generated action plans. Real-life leaders, on average, use the most number of pronouns (0.122)

[3]We report the exact API names, checkpoint dates for the evaluated LLMs, and the results from zero-shot prompting in the Appendix B. The results from three-shot prompting are reported in the main paper as they are directly comparable to the real-life leaders, as in both cases, three examples were provided.

[4]Our code and dataset can be found here: https://github.com/Vindhya-Singh/humansVsLLMs.git

and verbs (0.159), indicating a personal tone (Hynd and Chase, 1991) and action-oriented plans. LLMs such as Gemini-2.0-Flash (0.071), GPT-4o mini (0.070), Qwen-Plus (0.069), and Command-a-03-2025 (0.070) use more adjectives in their action plans than real-life leaders (0.051) and other LLMs, indicating more descriptive action plans.

In essence, real-life leaders are concise, action-oriented, and people-focused, indicating that they are better tuned to real-world action-planning and team engagement. LLMs vary dramatically in sentence length and words per sentence; therefore, they need to be adapted for the domain-specific audience. Overuse of nouns and underuse of verbs and pronouns across all LLMs make the action plans abstract or impersonal (Ehibor et al., 2025).

4.2 Who Writes More Readable Action Plans?

Analysis: We compared the readability (using the Flesch Reading Ease Score) (Kincaid et al., 1975; Tanprasert and Kauchak, 2021) and lexical diversity (Bestgen, 2025) of the LLM-generated action plans against those written by real-life leaders to determine textual accessibility and linguistic richness (Kriz et al., 2019; Kumar et al., 2020; Maddela et al., 2021) via TTR (Type-Token Ratio) and MATTR (Moving Average TTR) (Covington and McFall, 2010).

Results: Comparing the readability of action plans produced by LLMs and real-life leaders shows a stark contrast. Table 2 shows that real-life leaders write in the most readable way (Readability: 61.7, the highest among all, in the "standard" readable (60-70) range in the Flesch Reading Ease). All LLMs except for Llama-3.3-70b generate more readable action plans in the three-shot setting (see Tables 2 and 5). The exceptionally high readability scores of the action plans written by real-life leaders suggest their communication is easier to understand for broader audiences. It is noteworthy that the lexical diversity (TTR and MATTR) of the action plans by real-life leaders is higher than that of all the LLMs. This suggests nuanced and expressive language used by them, even though the LLMs have been trained on a wide corpus of online data. Moreover, LLM-generated action plans have readability scores below 35, except for Mistral-Large, indicating college-level or harder text ((30-50) in the Flesch Reading Ease). It is worthwhile to note that the lexical diversity (MATTR close to 0.99) of Gemini-2.0-Flash and Command-a-03-2025 is close to that of real-life leaders in the three-shot setting. Combined with the response-length and words per sentence results (Figure 3), this reflects complex sentence structures, technical vocabulary, and formal tones for most LLMs (Eder et al., 2023). However, some models like Llama-3.3-70b and Mistral-Large both have low MATTR ($M = 0.73$ and $M = 0.75$, respectively), indicating repetitive or limited vocabulary.

One of the most important implications of the above results is that the balance of clarity and diversity is rare in LLMs, as none match real-life leaders in balancing readability and high lexical diversity. Gemini-2.0-Flash (0.9992 ± 0.0011) performs well in MATTR in the three-shot setting, but it sacrifices readability (29.7630 ± 7.8788). Finally, some models like Llama-3.3-70b are low across metrics, indicating a struggle with both language variation and clarity. Overall, three-shot prompting has higher scores than zero-shot prompting (Tables 2 and 5 in Appendix B). Thus, we recommend using examples in prompts to improve performance and make better use of LLMs for domain-specific use cases.

4.3 How Similar Are Their Action Plans?

Analysis: Next, we evaluated the seven LLMs using BLEU (Papineni et al., 2001), ROUGE-L (Lin, 2004), and sentence cosine similarity scores (Zhang et al., 2020) to measure structural and lexical sentence overlap against leader-generated benchmarks.

Results: We found that in sentence similarity, Llama-3.3-70b ($M = 0.3479$) and Qwen-Plus ($M = 0.3365$) score the highest (Figure 4). This suggests that these two models produce the most semantically aligned action plans with real-life leaders. Gemini-2.0-Flash scored the lowest ($M = 0.261$), indicating weaker alignment with real-life leaders, despite decent ROUGE-L and BLEU scores. GPT-4o mini ($M = 0.117$) performs best in the ROUGE-L scores, indicating higher overlap of sequence (longest common subsequence) with real-life leaders' action plans. The other LLMs fall within a narrow band ($M = 0.108$ to 0.112), suggesting modest structural alignment. GPT-4o mini ($M = 0.0138$) and Mistral-Large ($M = 0.0135$) lead in BLEU scores, indicating better surface-level n-gram overlap. Command-a-03-2025 has the lowest BLEU score ($M = 0.0094$), which suggests more paraphrasing or different word usage. Across all three metrics (BLEU, ROUGE-L, Sentence Similarity), GPT-4o mini performs consistently well,

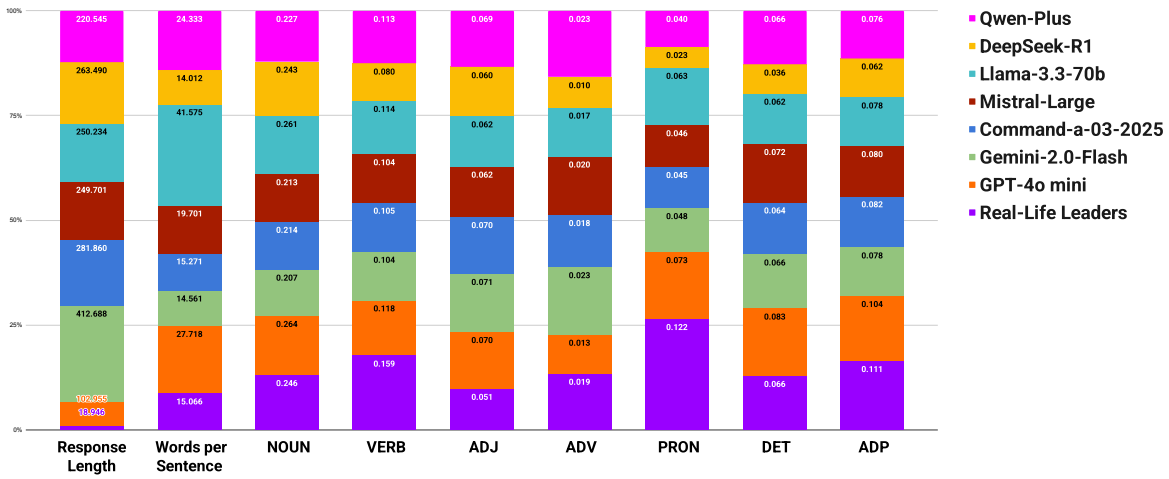

Figure 3: Analysis of structural variations in the action plans produced by real-life leaders and LLMs.

Source	Readability	TTR	MATTR
Real-Life Leaders	61.7412	0.9320	0.9998
GPT-4o-mini	32.1192 ± 12.3260	0.6565 ± 0.1286	0.9994 ± 0.0019
Gemini-2.0-Flash	29.7630 ± 7.8788	0.4925 ± 0.0430	0.9992 ± 0.0011
Command-a-03-2025	32.7575 ± 5.9299	0.5151 ± 0.0385	0.9996 ± 0.0007
Mistral-Large	35.0364 ± 8.1044	0.5049 ± 0.0495	0.7548 ± 0.0009
Llama-3.3-70b	17.7397 ± 21.0715	0.4840 ± 0.0462	0.7315 ± 0.0058
DeepSeek-R1	33.6825 ± 18.2526	0.6799 ± 0.1537	0.9868 ± 0.0564
Qwen-Plus	21.6531 ± 8.7629	0.6318 ± 0.0477	0.8812 ± 0.0009

Table 2: Readability and Lexical Diversity scores (MATTR) in three-shot settings.

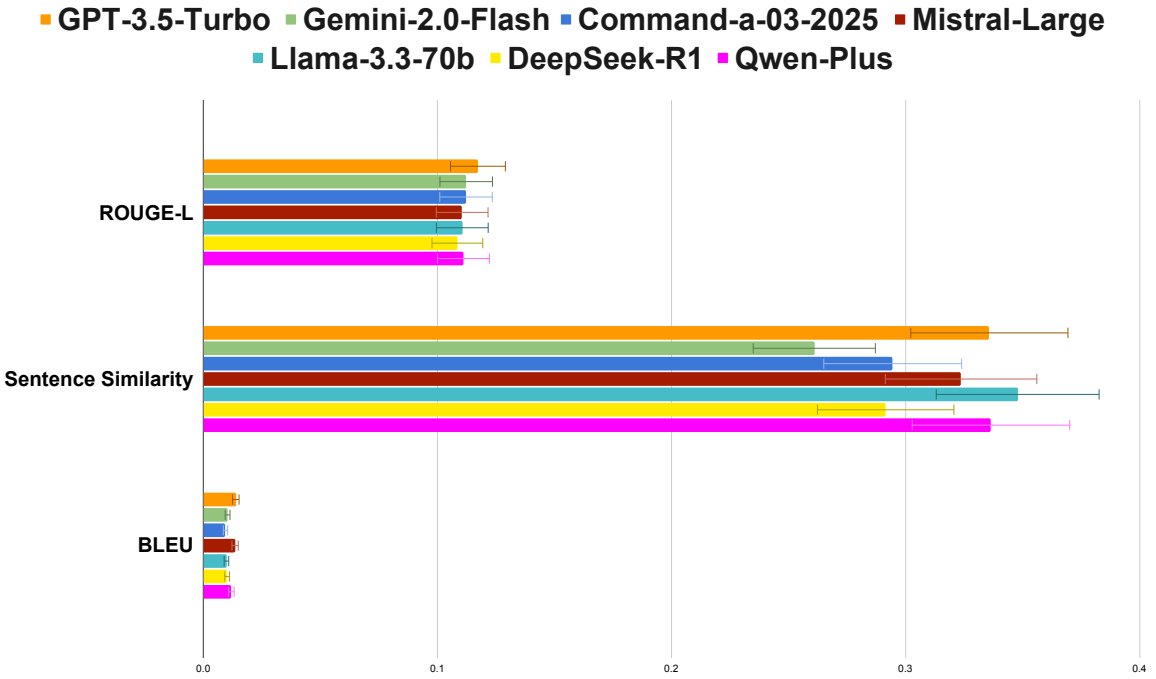

Figure 4: LLM evaluation: ROUGE-L, BLEU, and Sentence Similarity scores against human benchmarks.

indicating strong syntactic (BLEU, ROUGE-L) and decent semantic similarity with real-life leaders.

Even though Llama-3.3-70b performs the strongest in semantic similarity, it is weaker in other metrics (BLEU, ROUGE-L), indicating it is more meaning-aligned than word-aligned with real-life leaders. Notably, high BLEU/ROUGE-L scores do not always mean better understanding, as LLMs that use synonyms or paraphrasing may score lower but still be semantically correct.

4.4 Sentiment and Emotion Patterns in Action Plans

Analysis: We compare LLM-generated action plans to those of real-life leaders using 1) valence-arousal-dominance (Mohammad, 2025), and 2) the NRC Lexicon (Mohammad and Turney, 2010) to assess emotional and thematic alignment with established norms. Beyond linguistic analysis, analyzing the sentiments and emotional tone differences between LLMs and real-life leaders is a critical factor in leadership communication, where persuasion and emotional resonance matter (Uhl-Bien, 2006; A Rizvi and Sapna, 2021).

Results: Female ($M = 0.2525$) and Male Leaders ($M = 0.2499$) show comparable levels of valence, but less than most LLMs. LLMs like Llama-3.3-70b, GPT-4o mini, and Command-a-03-2025 are more assertive, positive, and composed than real-life leaders (see Figure 5). This is indicative of a design bias: models trained to be helpful may overly emphasize confidence and positivity (Steyvers et al., 2025). However, they can be used in simulated leadership scenarios as these models project idealized leadership personas who are calm, positive, and in control.

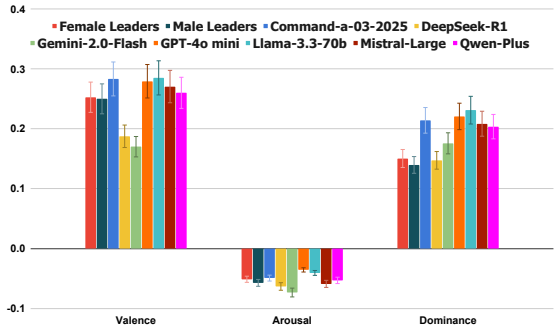

Figure 5: Valence-Arousal-Dominance scores.

Llama-3.3-70b ($M = 0.2309$), GPT-4o mini ($M = 0.2205$), and Command-a-03-2025 ($M = 0.2139$) express the highest sense of dominance, meaning their language conveys authority and direction. Female Leaders ($M = 0.1503$) and Male Leaders ($M = 0.1396$) are significantly lower in dominance, meaning real-life leaders tend to use less controlling, more egalitarian language (Hoogeboom et al., 2021). DeepSeek-R1 ($M = 0.147$) and Gemini-2.0-Flash ($M = 0.175$) are close to human levels in dominance, making statements more neutral or collaborative in tone. These results align with prior research (Chaves et al., 2019; Elsholz et al., 2019) stating that chatbots inherently aim to produce contextually appropriate language to increase end-user satisfaction, which is also reflected in our evaluation of LLMs.

Using the NRC Lexicon, we see in Table 3 that LLMs, especially GPT-4o mini, Mistral-Large, and Llama-3.3-70b, show high levels of trust (e.g., GPT-4o mini at 0.387). Qwen-Plus (0.37) and Gemini (0.365) show high positive emotions, more than any human group. Real-life leaders are less emotionally polarized and show more balance across categories, even though they use trust and positive terms frequently. Almost all LLMs exhibit lower sadness, fear, and disgust, suggesting their language is tuned for neutral to optimistic framing, except for DeepSeek-R1. It is interesting to note that female leaders use a broader emotional range than male leaders, as they have slightly higher scores in anger, sadness, and disgust, but use more surprise than their male counterparts, suggesting perhaps more dynamic or varied emotional framing. Notably, male leaders have higher trust (0.31) and anticipation (0.097) and significantly lower negative emotions than female leaders, suggesting that they favor a confident and strategic tone over emotional expressiveness (Lowenhaupt, 2021; de Vries et al., 2010). Therefore, with our diverse dataset, LLMs can be fine-tuned for inclusive and emotionally-variant communication style using the language of female leaders, while simultaneously, the language used by male leaders can facilitate fine-tuning LLMs for more strategic or formal tones. Finally, our comparative empirical analysis (Table 3) reveals that GPT-4o mini is the closest to Male Leaders, depicting high trust, low negativity, and high anticipation. Similarly, Mistral-Large is balanced, and closest to Female Leaders with high trust and joy, moderate surprise and sadness, that is, being emotionally expressive and positive. Interestingly, Qwen-Plus emerged rather optimistic, with high trust and joy, with some anger and disgust, indicating a balance between critique and positivity.

Source	F	Ang.	T	Surp.	P	N	Sad.	D	J	Ant.
Real-Life Leaders	0.0092	0.0034	0.3040	0.0352	0.2725	0.0251	0.0076	0.0027	0.0652	0.0947
Female Leaders	0.0096	0.0052	0.2977	0.0411	0.2738	0.0325	0.0115	0.0044	0.0610	0.0922
Male Leaders	0.0094	0.0030	0.3097	0.0348	0.2727	0.0209	0.0055	0.0020	0.0631	0.0975
Command-a-03-2025	0.0098	**0.0152**	0.3351	0.0323	0.3530	0.0257	0.0030	0.0104	0.1032	0.1123
DeepSeek-R1	**0.0360**	0.0110	0.2302	0.0276	0.3448	**0.0931**	**0.0469**	**0.0303**	0.0607	0.0863
Gemini-2.0-Flash	0.0103	0.0104	0.3127	0.0314	0.3654	0.0379	0.0055	0.0066	0.1033	0.1167
GPT-4o mini	0.0014	0.0038	**0.3870**	**0.0345**	0.3594	0.0143	0.0006	0.0008	0.0967	0.1015
Llama-3.3-70b	0.0047	0.0040	0.3600	0.0309	0.3430	0.0244	0.0036	0.0015	0.1077	**0.1202**
Mistral-Large	0.0054	0.0073	0.3692	0.0288	0.3377	0.0192	0.0028	0.0035	**0.1088**	0.1173
Qwen-Plus	0.0075	0.0110	0.3370	0.0262	**0.3770**	0.0190	0.0014	0.0086	0.0987	0.1135

Table 3: Emotion and sentiment distribution using the NRC Lexicon. F.: Fear, Ang.: Anger, T.: Trust, Surp.: Surprise, P: Positive, N.: Negative, Sad.: Sadness, D: Disgust, J: Joy, Ant.: Anticipation. GPT-4o mini aligns closest to Male Leaders, Mistral-Large best mirrors Female Leaders, while Qwen-Plus exhibits optimism with balanced critique. Highlighted values show the highest values across LLMs in each category.

4.5 Human Evaluation

Analysis: Comparing the action plans generated by LLMs to those of real-life leaders (Table 9) in terms of syntax, lexical diversity, and semantics did not capture their nuances, such as the quality of information. To that end, we used human evaluation to evaluate the action plans generated by LLMs and real-life leaders. Thus, we adopted the multi-dimensional evaluation framework proposed by Tam et al. (2024), and recruited human evaluators to assess Quality of Information (**accuracy, relevance, currency, comprehensiveness, coherence, usefulness**), Expression Style and Persona (**clarity, empathy**), Safety and Harm (**bias, fabrication**), and Trust and Confidence (**trust, satisfaction**). These dimensions were chosen to reflect real-world usability as leaders rely on trustworthy, well-articulated, and ethically sound advice, while safety ensures responsible LLM deployment. We randomly sampled more than 10 percent of action plans generated from each of the eight data sources (that is, seven LLMs and one from real-life leaders). We recruited 300 participants from Prolific, an online crowd-sourcing platform (Sieker et al., 2024), to evaluate the action plans against these twelve dimensions (highlighted above in bold). Out of all participants, 290 provided their demographic information. The participants were gender balanced (Male Evaluators=134, Female Evaluators=152, Trans-Female=2, Trans-Male=2) with an average age of 37.76 years. 73 percent of the evaluators held a bachelor's or master's degree. They were over 18 years old, fluent in English, were currently employed in a leadership position, and interacted with LLM-based chatbots at least once daily. Participants were well-instructed and signed a Consent Form upon which they were presented with eight action plans per question and were asked to rate them on a Best-Worst scale (Kiritchenko and Mohammad, 2017) as it is free from scale-bias, provides greater discrimination among items and between respondents, and gives better results with fewer respondents. We randomized the order of statements and questions to mitigate order bias, blinded raters to the source (real-life leaders vs. LLMs), and inserted a mid-survey attention check question. We excluded participants who failed the check from the evaluation.

Results: We compiled the results from our survey of 300 participants (for definitions and results, see Appendix B, Table 8, Figures 11 and 12) and found that action plans generated by Llama-3.3-70b were rated the best for *empathy*, better than Male and Female Leaders, suggesting that LLMs may simulate consistent emotional understanding. The action plans generated by Gemini-2.0-Flash were rated the most highly for *relevance*, *accuracy*, and *coherence*. It stands out as the most positively rated LLM, receiving the highest number of *Best* ratings and relatively few *Worst* ratings. This suggests consistent, high-quality, and reliable outputs across dimensions. Meanwhile, Qwen-Plus was the best rated for *satisfaction*. Gemini-2.0-Flash, GPT-4o mini, and Command-a-03-2025 performed well in *clarity*, *comprehensiveness*, *currency*, and *trust*. DeepSeek-R1 was the best-rated LLM for *usefulness*. DeepSeek-R1 and Command-a-03-2025 offer a more balanced profile with both strengths and weaknesses, as they appear frequently in both

USE CASE	BEST CHOICE	JUSTIFICATION	REMARKS
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	Most similar to real-life leader's action plans
Highest rated by human evaluators	Llama3.3-70b, Gemini-2.0-Flash, Command-a-03-2025, GPT-4o mini	Surveys	• Empathy: Llama stood out • Trust and Satisfaction: Gemini-2.0-Flash, Command-a-03-2025, GPT-4o mini performed well here, aligning with user confidence and response coherence.
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	Best mimic of Female Leader
Legal writing models/bots	DeepSeek-R1, GPT-4o mini	Best alignment in arousal and dominance, and NRC affects distribution	Best mimic of Male Leader
For AI leadership personas	GPT-4o mini, Mistral-Large, Qwen-Plus	Trustworthy, positive, engaging using NRC Lexicon	GPT-4o mini (best for Male Leader traits), Mistral-Large/Qwen-Plus (best for Female Leader traits)
Simulated leadership scenarios	GPT-4o mini, Llama-3.3-70b, Command-a-03-2025	High valence and dominance and highly rated by human evaluators	These LLMs feel more "leader-like" than real-life leaders
Balanced option	GPT-4o mini, Qwen-Plus, Gemini-2.0-Flash	Good blend of lexical, semantic, and human evaluation scores	Moderate emotional expressiveness of Gemini-2.0-Flash

Figure 6: Recommendations for future research use cases.

Best and *Worst* categories. Command-a-03-2025 and Llama-3.3-70b received high ratings for *bias*. These insights can inform future designs for LLM training and evaluations, especially when aiming to match human-like leadership communication skills.

5 Recommendations

LLMs are widely adopted tools in workplaces. In this interdisciplinary study, spanning NLP and Organizational Leadership, we draw on our findings and dataset to offer actionable insights and targeted recommendations for future research at the intersection of these domains. Future research should consider the following:

- To create inclusive and engaging communication tools, especially for workplaces, LLM fine-tuning should be based on real-life communication patterns, focusing on readability, simplification, and domain-specific language. These are the areas where our dataset can help.
- Improving inclusion frameworks may require acknowledging where LLMs outperform human leaders, especially in emotional consistency and coherence.
- For stronger practical utility, future work should also assess action plans against each SMART criterion in greater detail.

We also provide a use-case-specific summary of our recommendations in Figure 6.

6 Conclusion

Our work advances research on LLMs in high-stakes professional settings such as leadership. We collected a diverse dataset from over 250 real-life leaders, providing a valuable resource for future studies. Our comparative analysis shows that current LLMs often diverge from authentic leadership communication in readability, intensity, and authority. By grounding fine-tuning in real-world communication patterns, emphasizing readability, simplification, and domain-specific language, future models can become more inclusive and effective in workplace applications. Our dataset enables researchers and developers to adapt LLM behavior to leadership contexts and communication goals. We envision it as a springboard for responsible, context-aware LLM deployment in domains such as organizational leadership and human resource management.

Limitations

The primary limitation of our work is that we evaluate and compare only texts in English. As leadership is global, we aim to develop our Leader Success Bot further for leaders who interact with their team members in languages other than English. To adapt our script in multiple languages, we plan to collaborate with multilingual domain experts. Second, even though approximately 32 leaders in our dataset reported disabilities, we need to include more leaders with disabilities and adapt our chatbot for greater accessibility, such as leaders who are blind. In its present version, our Leader Success Bot is not designed for blind users. Third, although we evaluated a diverse set of chatbot-based LLMs, their swift and competitive development means that benchmarking them against real-life leaders remains an iterative, scalable process. Finally, we compare action plans generated by the LLMs to those of real-life leaders, on the basis of gender, as this aligns with established research in organizational leadership (Archer and Kam, 2022) and allows for straightforward comparison. However, our dataset also enables exploration of other intersecting identities, such as age or leadership experience, which could further expand the literature.

Ethical Considerations

We included leaders from a diverse range of ethnic backgrounds, age groups, leadership experiences, and physical abilities. The Institutional Review Board (IRB Protocol Number: HSR-24-25-53) approved the data collection part of our research at California State University, Fullerton. The Ethical Council of the Max Planck Society (Protocol Number: 2021_29) approved the human evaluation part of our research. We adhere to best research practices to ensure participant confidentiality. Following Jernite and colleagues' (Jernite et al., 2022) ethical framework, our research prioritized harm reduction by securing leaders' and human evaluators' informed consent, anonymizing all identifying details, and maintaining clear communication about data usage and potential risks involved. We do not store the names or affiliations of the participants. The data from the participants is anonymized using a randomly generated unique identifier, with only their interaction with the bot and sociodemographic details retained. The participants were well informed about the study, and we used their data after receiving their informed consent. The participants who interacted with our Leader Success Bot and the human evaluators were paid for their time.

Acknowledgments

We are grateful to Dr. H. Phoenix Van Wagoner of California State University, Fullerton, for his expert guidance on chatbot script development and support throughout the Institutional Review Board (IRB) proposal process. His contributions were crucial to the successful completion of this study.

We are thankful to the SemRel group at the University of Stuttgart for their feedback.

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting our work.

References

Irfan A Rizvi and Popli Sapna. 2021. Revisiting leadership communication: A need for conversation. *Global Business Review*, 0(0):09721509211061979.

Cloud Alibaba. 2025. Tongyi qianwen (qwen).

Allison M.N. Archer and Cindy D. Kam. 2022. She is the chair(man): Gender, language, and leadership. *The Leadership Quarterly*, 33(6):101610.

Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*, pages 52–61, New Brunswick, NJ, USA. ACM Press.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Yves Bestgen. 2025. Estimating lexical diversity using the moving average type-token ratio (MATTR): Pros and cons. *Research Methods in Applied Linguistics*, 4(1):100168.

Janet A. Boekhorst. 2015. The role of authentic leadership in fostering workplace inclusion: A social information processing perspective. *Human Resource Management*, 54(2):241–264.

Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2024. How knowledge workers use and want to use LLMs in an enterprise context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, Honolulu HI USA. ACM.

Abraham Carmeli, Roni Reiter-Palmon, and Enbal Ziv. 2010. Inclusive leadership and employee involvement in creative tasks in the workplace: The mediating role of psychological safety. *Creativity Research Journal*, 22(3):250–260.

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109, Kyoto Japan. ACM.

Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. 2024. Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, 133:103002.

Andrei-Victor Chisca, Andrei-Cristian Rad, and Camelia Lemnaru. 2024. Prompting fairness: Learning prompts for debiasing Large Language Models. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 52–62, St. Julian's, Malta. Association for Computational Linguistics.

Alexander Choi, Syeda Sabrina Akter, J.P. Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Suk Bong Choi, Thi Bich Hanh Tran, and Byung Il Park. 2015. Inclusive leadership and work engagement: Mediating roles of affective organizational commitment and creativity. *Social Behavior and Personality: An international journal*, 43(6):931–943.

Team Cohere. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Niamh E. A. Dawson, Stacey L. Parker, and Tyler G. Okimoto. 2024. Profiles of diversity and inclusion motivation: Toward an employee-centered understanding of why employees put effort into inclusion and exclusion. *Human Resource Management*, 63(1):45–66.

Reinout de Vries, Angelique Bakker-Pieper, and Wyneke Oostenveld. 2010. Leadership = Communication? The relations of leaders' communication styles with leadership styles, knowledge sharing and leadership outcomes. *Journal of Business and Psychology*, 25:367–380.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

David Dwertmann and Stephan Boehm. 2016. Status matters: The asymmetric effects of supervisor–subordinate disability incongruence and climate for inclusion. *The Academy of Management Journal*, 59.

Elisabeth Eder, Ulrike Krieg-Holz, and Michael Wiegand. 2023. A question of style: A dataset for analyzing formality on different levels. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 580–593, Dubrovnik, Croatia. Association for Computational Linguistics.

Oremire J. Ehibor, Joy Eyisi Jr, Jonathan A. Odukoya, Charles U. Ogbulogo, C. U. C. Ugorji, Onyekachi Odo, Lily Chimuanya, Eugenia Abiodun-Eniayekan, Edith Awogu-Maduagwu, and Rebecca U. Adesiyan. 2025. Linguistic-stylistic analysis of the language of leadership in the political arena and the business world. *Cogent Arts & Humanities*, 12(1):2464382.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305, Glasgow Scotland UK. ACM.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in Large Language Models: A survey. *Computational Linguistics*, 50(3):1097–1179.

DeepMind Google. 2024. Gemini 2.0 Flash.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of GPT-3. *Preprint*, arXiv:2209.12356.

Jennifer P. Green, Reeshad S. Dalal, Shea Fyffe, Stephen J. Zaccaro, Dan J. Putka, and David M. Wallace. 2023. An empirical taxonomy of leadership situations: Development, validation, and implications for the science and practice of leadership. *Journal of Applied Psychology*, 108(9):1515–1539.

Yves R.F. Guillaume, Jeremy F. Dawson, Lilian Otaye-Ebede, Stephen A. Woods, and Michael A. West. 2017. Harnessing demographic differences in organizations: What moderates the effects of workplace diversity? *Journal of Organizational Behavior*, 38(2):276–303.

Fabrice Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. *Preprint*, arXiv:2406.12680.

Jeongyun Heo and Jiyoon Lee. 2019. Cisa: An inclusive chatbot service for international students and academics. In *HCI International 2019 – Late Breaking Papers: 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings*, page 153–167, Berlin, Heidelberg. Springer-Verlag.

Jenny M. Hoobler and Daniel J. Brass. 2006. Abusive supervision and family undermining as displaced aggression. *Journal of Applied Psychology*, 91(5):1125–1133.

Marcella A.M.G. Hoogeboom, Aaqib Saeed, Matthijs L. Noordzij, and Celeste P.M. Wilderom. 2021. Physiological arousal variability accompanying relations-oriented behaviors of effective leaders: Triangulating skin conductance, video-based behavior coding and perceived effectiveness. *The Leadership Quarterly*, 32(6):101493.

Cynthia R. Hynd and Nancy D. Chase. 1991. The relation between text type, tone, and written response. *Journal of Reading Behavior*, 23(3):281–306.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2206–2222, New York, NY, USA. Association for Computing Machinery.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. *Institute for Simulation and Training*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst scaling more reliable than rating scales: A case study on sentiment inntensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Sami Koivunen, Saara Ala-Luopa, Thomas Olsson, and Arja Haapakorpi. 2022. The march of chatbots into recruitment: Recruiters' experiences, expectations, and design opportunities. *Computer Supported Cooperative Work (CSCW)*, 31(3):487–516.

Ayfer Veli Korkmaz, Marloes L. Van Engen, Lena Knappert, and René Schalk. 2022. About and beyond leading uniqueness and belongingness: A systematic review of inclusive leadership research. *Human Resource Management Review*, 32(4):100894.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North*, pages 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Christina N. Lacerenza, Denise L. Reyes, Shannon L. Marlow, Dana L. Joseph, and Eduardo Salas. 2017. Leadership training design, delivery, and implementation: A meta-analysis. *Journal of Applied Psychology*, 102(12):1686–1718.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705–717.

Rebecca Lowenhaupt. 2021. *The structure of leadership language: Rhetorical and linguistic methods for studying school improvement*, pages 137–153. Springer International Publishing, Cham.

Daniela Macari, Alex Fratzl, Ksenia Keplinger, and Christoph Keplinger. 2024. Accelerating the pace of innovation in robotics by fostering diversity and inclusive leadership. *Science Robotics*, 9(97):eadt1958.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Hannah Mayer, Lareina Yee, Michael Chui, and Roger Roberts. 2025. Superagency in the workplace: Empowering people to unlock AI's full potential.

Akshay Mendhakar and Darshan H S. 2024. Parts-of-Speech (PoS) analysis and classification of various text genres. *Corpus-based Studies across Humanities*, 1(1):99–131.

Meta. 2024. Llama 3.3: Model cards and prompt formats.

Mistral. 2024. Au Large: Mistral AI.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad. 2025. NRC VAD Lexicon v2: Norms for valence, arousal, and dominance for over 55k English terms. *Preprint*, arXiv:2503.23547.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10).

Mehran Nasseri, Patrick Brandtner, Robert Zimmermann, Taha Falatouri, Farzaneh Darbanian, and Tobechi Obinwanne. 2023. Applications of large language models (LLMs) in business analytics – Exemplary use cases in data preparation tasks. In Helmut Degen, Stavroula Ntoa, and Abbas Moallem, editors, *HCI International 2023 – Late Breaking Papers*, volume 14059, pages 182–198. Springer Nature Switzerland, Cham.

Ingrid M. Nembhard and Amy C. Edmondson. 2006. Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behavior*, 27(7):941–966.

Toby Newstead, Bronwyn Eager, and Suze Wilson. 2023. How AI can perpetuate – Or help mitigate – Gender bias in leadership. *Organizational Dynamics*, 52(4):100998.

Lisa H. Nishii. 2013. The benefits of climate for inclusion for gender-diverse groups. *Academy of Management Journal*, 56(6):1754–1774.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence.

Aneesya Panicker, Rakesh Kumar Agrawal, and Utkal Khandelwal. 2018. Inclusive workplace and organizational citizenship behavior: Study of a higher education institution, India. *Equality, Diversity and Inclusion: An International Journal*, 37(6):530–550.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Amy E. Randel, Benjamin M. Galvin, Lynn M. Shore, Karen Holcombe Ehrhart, Beth G. Chung, Michelle A. Dean, and Uma Kedharnath. 2018. Inclusive leadership: Realizing positive outcomes through belongingness and being valued for uniqueness. *Human Resource Management Review*, 28(2):190–203.

McKenzie Raub. 2018. Bots, bias and big data: Artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Ark. L. Rev.*, 71:529.

Lynn M. Shore, Amy E. Randel, Beth G. Chung, Michelle A. Dean, Karen Holcombe Ehrhart, and Gangaram Singh. 2011. Inclusion and diversity in work groups: A review and model for future research. *Journal of Management*, 37(4):1262–1289.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The illusion of competence: Evaluating the effect of explanations on users' mental models of visual question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19459–19475, Miami, Florida, USA. Association for Computational Linguistics.

Geovana Ramos Sousa Silva and Edna Dias Canedo. 2024. Towards user-centric guidelines for chatbot conversational design. *International Journal of Human–Computer Interaction*, 40(2):98–120.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Mary Uhl-Bien. 2006. Relational leadership theory: Exploring the social processes of leadership and organizing. *The Leadership Quarterly*, 17:654–676.

Daan Van Knippenberg, Lisa H. Nishii, and David J. G. Dwertmann. 2020. Synergy from diversity: Managing team diversity to enhance performance. *Behavioral Science & Policy*, 6(1):75–92.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Nature Scientific Reports*, 13.

Ned Wellman. 2017. Authority or community? A relational models theory of group-level leadership emergence. *Academy of Management Review*, 42(4):596–617.

Marta Witkowska, Joanna Dołżycka, Caterina Suitner, and Magdalena Formanowicz. 2024. The grammar of persuasion: A meta-analytic review disconfirming the role of nouns as linguistic cues of subsequent behavior. *Journal of Language and Social Psychology*, 43(4):428–449.

Ziang Xiao, Michelle X. Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 437–447, Marina del Ray California. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for news summarization. *Preprint*, arXiv:2301.13848.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Appendix A

A.1 Chatbot Design-Additional Information on the Design and Development of the Leader Success Bot

The Leader Success Bot guides leaders to set daily inclusion action plans. These inclusion action plans are supported by examples showcasing leader inclusion behavior, such as Recognizing Uniqueness, Showing Appreciation, Strengthening Belongingness, and Supporting Organizational Efforts. Building on the existing research (Silva and Canedo, 2024), our bot uses state-of-the-art messaging technology to integrate validated measures of affects and structured planning of inclusion behaviors within engaging, user-centric interactions that include empathic responses, emojis, guided examples, and information on demand. Participants experience a fully interactive environment through buttons, images, emojis, and questionnaires, promoting psychological engagement that is especially beneficial in long-term interventions. The front end of the bot focuses on two main platforms for user interaction and data collection: Qualtrics and Telegram. These platforms are chosen for their large and diverse user bases, allowing for robustness, efficient interaction, and handling large-scale surveys with thousands of participants without performance issues. While Prolific is used to recruit

participants, Qualtrics is used to collect data from surveys (administered through the chatbot). The script-based Telegram chatbot is used for real-time conversational interaction (action-planning and reflection) and administering the Qualtrics surveys during various stages of the experiment. The backend infrastructure is built on a robust and scalable stack, ensuring smooth operation and accurate data collection. The core components include a MongoDB database (for storing user interactions with the chatbot) and a Heroku server setup (for hosting the chatbot application), which are optimized for handling the experimental data and user interactions. Our Leader Success Bot is programmed using JavaScript and JSON, facilitating automated interaction on Telegram (see Figure 7).

To participate, leaders were required to connect to the Leader Success Bot for two work weeks, develop inclusion action plans in the morning, reflect on their emotions and action-plan progress in the evening (both at the time of their own choosing) (see Figure 8). The chatbot guided leaders through the process of setting SMART inclusion actions for each inclusion dimension (showing appreciation, recognizing uniqueness, strengthening belonging, and supporting organizational efforts), making each action specific, measurable, achievable, relevant, and time-bound. During the interaction, the Leader Success Bot provided detailed examples and offered on-demand information to explain each inclusion dimension in depth, supporting leaders in developing specific, observable inclusion behaviors. In the first week, the bot guides leaders to set daily inclusion action plans for recognizing uniqueness and showing appreciation to their team members. Starting with the second week of the experiment, the bot guides the leaders to set inclusion action plans for strengthening belongingness and supporting organizational efforts.

B Appendix B

B.1 API Names and Checkpoint Dates for the Evaluated LLMs

Table 4 details the exact API names and checkpoint dates for the evaluated LLMs.

LLM Names	API Names	Month
Command-a-03-2025	Cohere API	April
DeepSeek-R1	DeepSeek API	April
Gemini-2.0-Flash	Gemini API	April
GPT-4o mini	OpenAI API	April
Llama-3.3-70b	LLM API	April
Mistral-Large	Mistral API	April
Qwen-Plus	Alibaba API	April

Table 4: API information of the evaluated LLMs. All the LLMs were prompted in the year 2025.

B.2 Zero-Shot Prompting: Comparing LLM-generated action plans with Real-life Leaders

The analysis of structural variations reveals significant differences between the action plans of real-life leaders and various LLMs (See Figure 9). Real-life leaders exhibit notably shorter responses and lower words-per-sentence metrics, suggesting a more concise communication style. Their language is marked by a higher frequency of verbs and pronouns, which indicates they focus on action and interpersonal engagement. In contrast, LLMs such as GPT-4o mini, Gemini-2.0-Flash, and others tend to produce much longer responses with more complex sentence structures. These models also favor nouns and determiners over verbs and pronouns, indicating a more descriptive and less dynamic style. Among the LLMs, Qwen-Plus and GPT-4o mini stand out for their relatively high use of verbs, aligning them closer to human speech patterns. Real-life leaders demonstrate the highest readability by a wide margin, with a score close to 62, significantly surpassing all LLMs. Among the models, DeepSeek-R1 and Llama-3.3-70b come closest in readability but still fall far short. In terms of lexical diversity, as measured by TTR and MATTR, real-life leaders again lead, particularly on MATTR, where most models perform well except Mistral-Large and Llama-3.3-70b. Gemini-2.0-Flash and Command-a-03-2025 show exceptionally high MATTR scores, nearly matching real leaders, but their TTR and readability remain low. Finally, while LLMs are more verbose and structurally complex, real-life leaders prioritize readability (see Table 5).

B.3 Examples of Prompts and Generated Responses

Example prompts used for three-shot and zero-shot sociodemographic prompting of the seven LLMs are depicted in Figure 10. We used the age, ethnic background, gender, and leadership experience of the leaders, along with the script designed with domain experts, to prompt the LLMs. This provided the LLMs with a persona and a detailed description

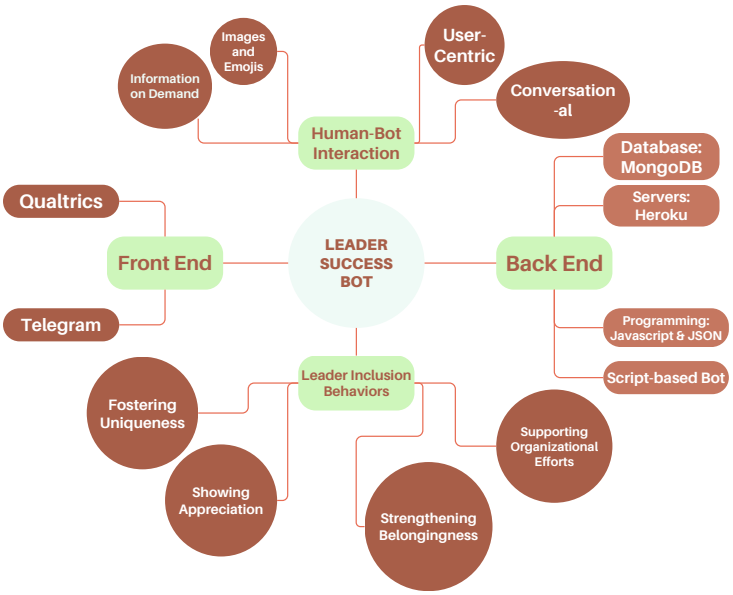

Figure 7: Design and implementation of the Leader Success Bot.

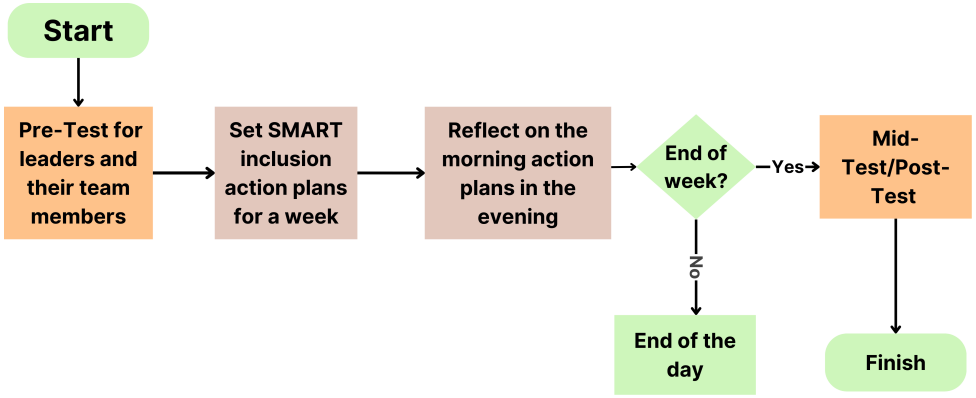

Figure 8: Interaction procedure followed by the Leader Success Bot.

Source	R	TTR	M
Real-Life Leaders	61.7412	0.9320	0.9998
Gemini-2.0-Flash	27.7299	0.4702	0.9990
GPT-4o mini	25.4071	0.5531	0.8580
Mistral-Large	30.9190	0.4835	0.8654
Llama-3.3-70b	31.0956	0.4415	0.8436
Qwen-Plus	20.3764	0.6309	0.9237
Command-a-03-2025	29.5644	0.4923	0.9922
DeepSeek-R1	35.8840	0.5300	0.9547

Table 5: Readability and Lexical Diversity scores in zero-shot settings. R: Readability, M: MATTR.

of the task. As inclusion is an abstract concept, for the three-shot prompting, we provided examples on setting SMART inclusion action plans, while for the zero-shot prompting, there was only an instruction and no examples.

B.4 Comparing words used in LLM-generated action plans with Real-life Leaders for Three-Shot Prompting

Additionally, we extracted the top 20 Nouns and Verbs (Green et al., 2023; Witkowska et al., 2024) used by LLMs and real-life leaders in their action plans and compared them. As these are action plans, identifying the verbs used correlates with the actions and interactions of leaders with their teams. Female and Non-Binary leaders lean more toward inclusive, emotionally intelligent language, while Male Leaders balance goals and collaboration. Overuse of "ensure", "implement", and "foster" in nearly every LLM suggests safe but sterile leadership communication, lacking situational variability (Lowenhaupt, 2021; de Vries et al., 2010; Hoogeboom et al., 2021; Eder et al., 2023). These results are summarized in Table 6.

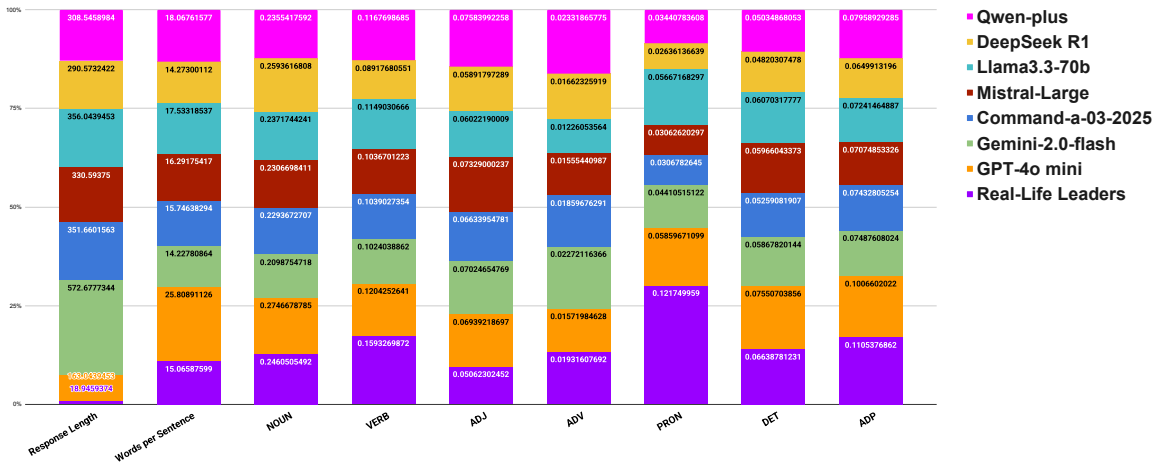

Figure 9: Analysis of structural variations in action plans in zero-shot settings.

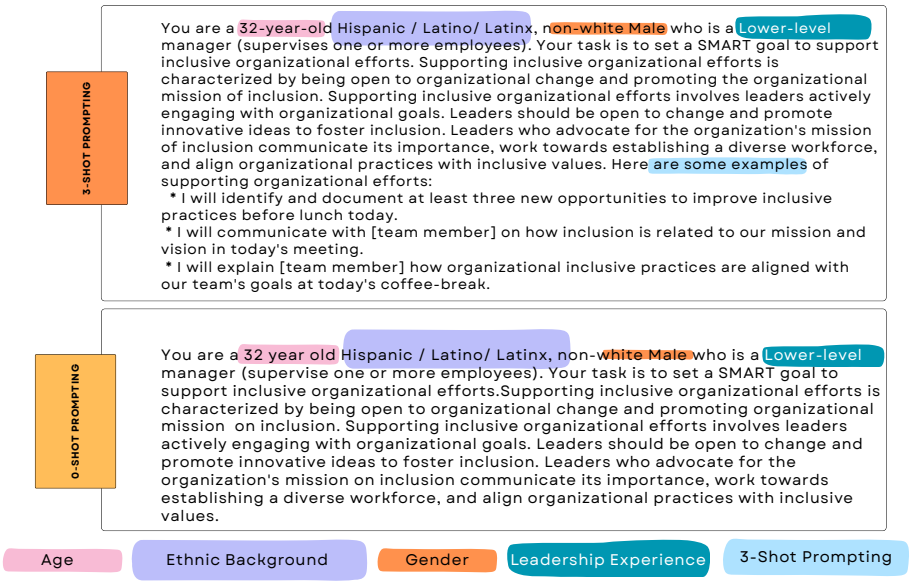

Figure 10: Example prompts for the three-shot and zero-shot sociodemographic prompting.

In our analysis, we extracted all nouns and verbs used by real-life leaders and the seven LLMs used in this paper. We found that while words such as "team", "member", "meeting", "ensure", and "meet" are used by both real-life leaders and LLMs, some words (nouns and verbs) are used extensively by LLMs but never by any real-life leader (see Table 7).

B.5 LLM Recommendations for Specific Use-Cases

Based on our analysis in the paper, we have summarized our recommendations in Figure 6. Our recommendations come from the empirical evidence collected for this research.

B.6 Human Evaluation

We evaluated the action plans generated by LLMs and written by real-life leaders on twelve dimensions (Tam et al., 2024), as tabulated in Table 8. The results are visualized in the Figures 11 and 12. Example action plans are tabulated in Table 9.

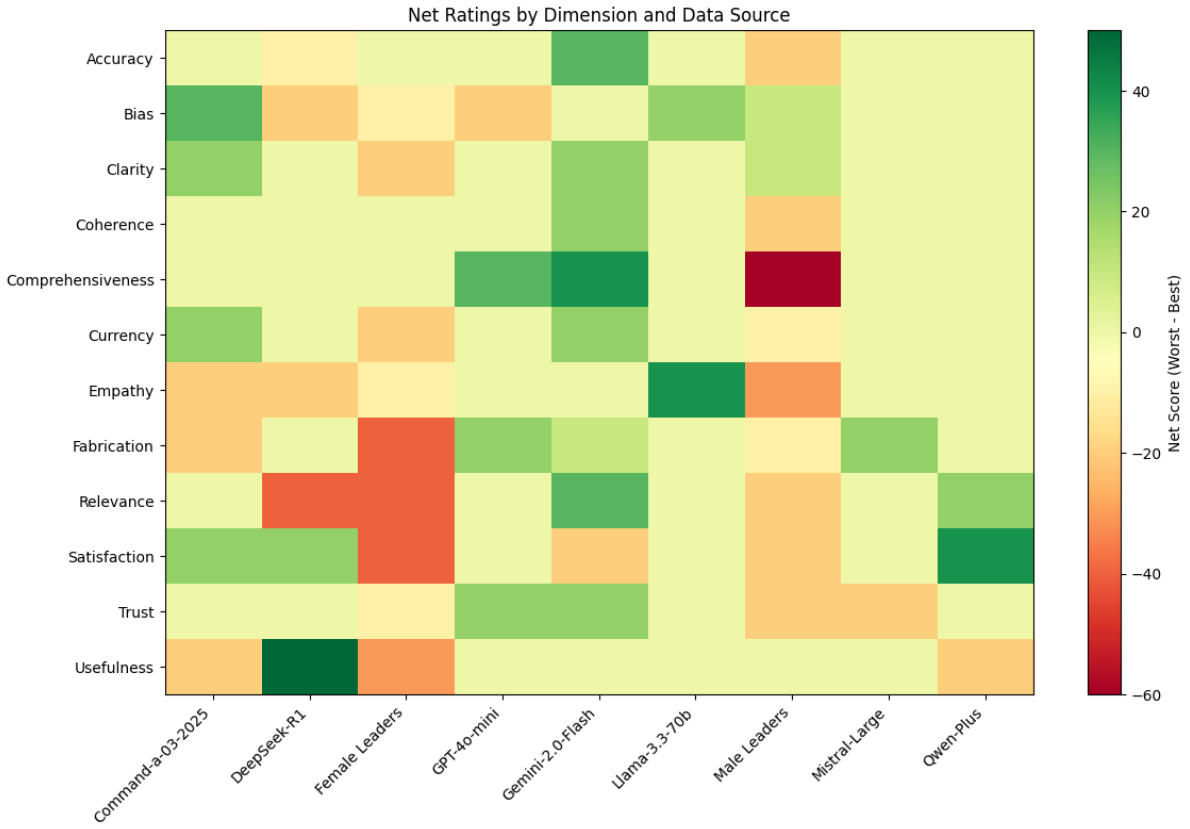

Figure 11: Human evaluation results reported using Best–Worst scaling. Items were presented in randomized order, and raters were blinded to the source (real-life leaders vs. LLMs).

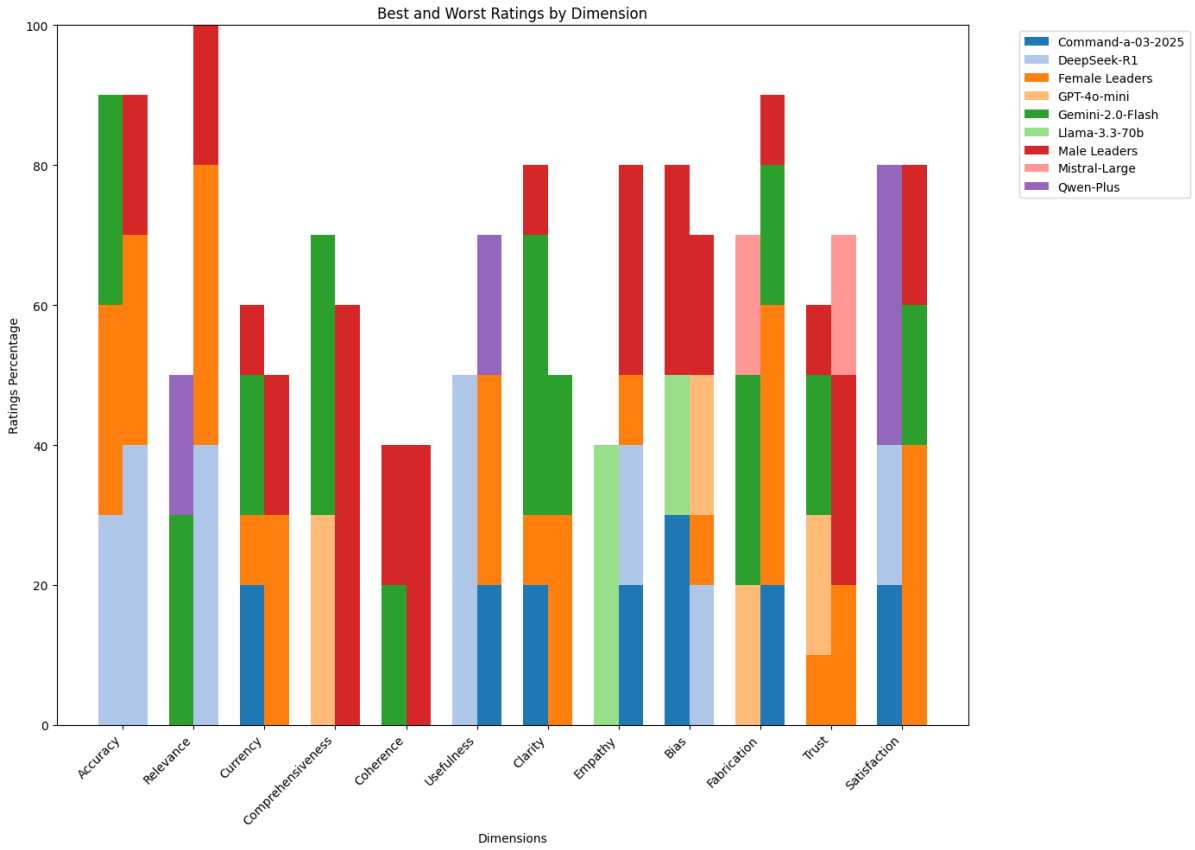

Figure 12: Visualization of the survey ratings.

Group	Overall Tone	Focus of Action: Top Verbs	Leadership Style
LLMs	Structured, procedural	Ensuring, fostering, implementing, creating	Strategic, formal, supportive
Female Leaders	Relational, empathetic	Ensuring, offering, helping, praising	Supportive, nurturing, people-focused
Male Leaders	Balanced, pragmatic	Working, improving, giving, discussing	Goal-oriented, collaborative
Non-Binary Leaders	Contextual, emotional	Dealing, decompressing, organizing	Adaptive, emotionally intelligent

Table 6: Mapping leadership styles of various groups based on the top verbs used by them in their action plans.

Nouns never used by real-life leaders	Verbs never used by real-life leaders
action	foster
activities	ensure
appreciation	identify
belongingness	provide
billing	align
breakdown	allocate
categories	support
contributions	lead
details	recognize
development	complete
diversity	create
efforts	contribute
environment	track
equity	implement
ideas	share
inclusion	empower
information	promote
meeting workshop	support
mission	let
month	rotate
number	recognize
participation	implement
plan	include
program	work
progress	strengthen
quarter	provide
quota	measure
relevant	focus
September	outline
session	strengthen
skills	organize
smart	promote
specific	boost
strengths	achieve
system	aim
type	incorporate
weeks	learn

Table 7: List of Nouns and Verbs used only by LLMs but never by any real-life leader.

Principle	Dimension	Definition
Quality of Information	Accuracy	Correctness of response provided by the LLM.
	Relevance	Alignment of response provided by the LLM to the user's query. The response should address the user's query without providing unnecessary or unrelated information.
	Currency	The response should contain the most current knowledge available, especially if the topic is one where new data or developments frequently occur.
	Comprehensiveness	Completeness of response provided by the LLM. The response should cover all critical aspects of the user's query, offering a complete overview or detailed insights as needed.
	Coherence	Coherence of response with established facts and theories. The response should not contradict itself.
	Usefulness	Applicability and utility of the response. The response should be of practical value, actionable, and applicable to the user's context or problem.
Expression style and persona	Clarity	Quality of the response is clear, understandable, and straightforward, making it easy for the user to comprehend the provided response.
	Empathy	Ability of the LLM to generate a response that recognizes and reflects the emotions or tone conveyed in the user's input, simulating a considerate and understanding interaction.
Safety and harm	Bias	Presence of systematic prejudices in the response, such as racial or gender bias.
	Fabrication	Fabrication is when the response contains entirely made-up information or data and includes plausible but non-existent facts.
Trust and Confidence	Trust	Confidence in the LLM that it will provide accurate, fair, and safe responses.
	Satisfaction	The LLM meets or exceeds the expectations of the user in terms of response quality and relevance.

Table 8: Evaluation dimensions and their definitions.

Source	Example Action Plans
Real-Life Leaders	I will meet with [team member] to discuss outstanding training and set her a goal of completing 50% of training by the end of October
GPT-4o mini	By the end of this month, I will organize and facilitate a monthly team lunch where team members can openly share their thoughts, experiences, and ideas to strengthen belongingness within the team.
Gemini-2.0-Flash	I will implement a "Team Kudos" board within one week where team members can publicly acknowledge each other's contributions, with at least 3 separate entries of positive contributions documented each week for one month. I will allocate 30 minutes each week to solicit and record at least one "Kudos" entry per day to ensure success.
Command-a-03-2025	I will personally recognize and appreciate the efforts of at least two team members each week by highlighting their specific contributions during our weekly team meetings or through written acknowledgments (e.g., emails or Slack messages).
Mistral-Large	Starting from the beginning of next month, I will schedule and conduct bi-weekly 30-minute one-on-one meetings with each team member to understand their unique perspectives, address any challenges they might be facing, and discuss their ideas to promote belongingness within the team.
Llama-3.3-70b	Within the next 6 weeks, I will schedule monthly one-on-one check-ins with each of my team members to build stronger relationships and ensure everyone's voice is heard. During these check-ins, I will ask for their input and feedback on current projects and team decisions, and provide opportunities for them to share their ideas and concerns. I will also make sure to follow up on their suggestions and ideas, and provide explanations for the decisions made by the team. Additionally, I will volunteer to help organize a quarterly team-building event, starting with a social gathering within the next 3 months, to foster positive interactions and a sense of community among team members.
DeepSeek-R1	By the end of each month for the next three months, I will strengthen belongingness within my team by initiating a structured, anonymous feedback survey to gather input on team dynamics and equity, then organize one team-building activity based on the consensus of suggestions. I will ensure 100% participation by sending reminders and offering flexible response times, and I will transparently share how feedback shaped the activity in a follow-up email.
Qwen-Plus	By the end of this month, I will meet with each team member individually for 30 minutes to discuss their personal career aspirations, identify one unique strength they bring to the team, and collaboratively create a plan to enhance that strength through specific training or project opportunities.

Table 9: Example action plans generated by real-life leaders and various LLMs.