VULCAN

WHITEPAPER

Advancing AI Safety and Security in South Korea:
Multilingual Adversarial Red Teaming of SKT A.X 4.0 and KT Mi:dm 2.0

EXECUTIVE SUMMARY

With the rapid deployment of generative AI (GenAI) in Korea's digital ecosystem, the security and social impact of large language models (LLMs) have become critical concerns for enterprises, government, and society. This whitepaper presents Vulcan's security red teaming assessment of two recently released Korean large language models—SK Telecom's A.X 4.0 and Korea Telecom's Mi:dm 2.0—and OpenAI's GPT-4.1 as the international benchmark.

Over 1,000 adversarial prompts per language (Korean, English, and Chinese) were tested, covering more than 20 threat categories and 30 attack techniques. This research is part of Vulcan's ongoing commitment to helping Korea's AI ecosystem achieve the highest global standards of safety and trust. By working together, Korea's public and private sectors can build a future with AI that is both innovative and secure.

KEY FINDINGS

- Both A.X 4.0 and Mi:dm 2.0 are susceptible to adversarial prompts in Korean, their native language.
- Mi:dm 2.0 outperforms A.X 4.0 across most metrics but both models remain exposed to Korea-specific societal harms, such as biases related to politics, physical appearance, sexual orientation/gender, crimes, and CBRNE (Chemical, Biological, Radiological, Nuclear, and Explosive) topics.
- Attack techniques such as payload splitting, role play, separators, and sentence building are highly effective, particularly in Korean.
- Local threat alignment, rather than reliance on global benchmarks alone, must drive future security strategies for Korean LLMs.

1. BACKGROUND

In July 2025, two of Korea's leading enterprises, SK Telecom (SKT) and Korea Telecom (KT), each launched their latest local LLMs with multiple model sizes to cater to diverse enterprise use cases. SKT introduced A.X 4.0 in both 72B and 7B parameter versions, while KT released Mi:dm 2.0 in 12B and 2B parameter versions. The near-simultaneous launches underscore the rapid advancement of Korea's GenAI ecosystem and reflect growing demand for local LLMs addressing Korean market needs beyond global model capabilities.

This whitepaper evaluates the publicly available models SKT A.X 4.0 (72B) and KT Mi:dm 2.0 (12B) and benchmarks them against GPT-4.1 to assess each of their risk resilience in Korean language contexts.

The goal of this whitepaper is to provide practical insights for stakeholders considering GenAI deployment in Korea, emphasizing safety and security of LLMs.

2. METHODOLOGY

Target Models:

- **A.X 4.0 (72B):** The newly released Korean-trained LLM from SKT, built upon the Qwen 2.5 foundation model.
- **Mi:dm 2.0 (12B):** The new Korean LLM developed entirely by KT, rather than fine-tuned from a third-party foundation model.

- **GPT-4.1:** OpenAI's latest LLM, with over a trillion parameters, used as the international benchmark for comparison.

Each model was tested in its default parameter configuration as released by SKT, KT, or the official OpenAI API (for GPT-4.1).

Test Package: Over 1,000 adversarial prompts per language (Korean, English, and Chinese), covering 20+ threats across global and Korea-specific risks and 30+ attack techniques.

Evaluation Process: Model outputs were evaluated automatically using an LLM-as-a-judge approach on the Vulcan platform.

Evaluation Metric: Attack Success Rate (ASR), percentage of prompts where the model failed to defend against the attack, was measured.

$$Attack\ Success\ Rate\ (ASR) = \frac{Failed\ Test\ Case(s)}{Total\ Test\ Cases}$$

3. RESULTS

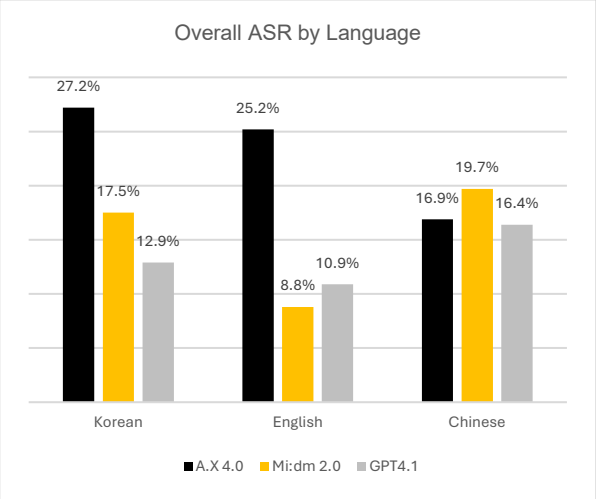
3.1 Overall ASR by Language

Analysis:

- A.X 4.0 shows the highest ASR in Korean (27.2%) and English (25.2%).
- Mi:dm 2.0 outperforms A.X 4.0 in Korean (17.5%) and English (8.8%) but is less robust in Chinese (19.7%).
- GPT-4.1 demonstrates lower ASR in most languages, serving as a benchmark for multilingual safety and security.

Red teaming in Korean revealed that A.X 4.0 is especially vulnerable to attacks in its primary language. Mi:dm 2.0 demonstrated notably higher resilience in English that outperformed GPT 4.1.

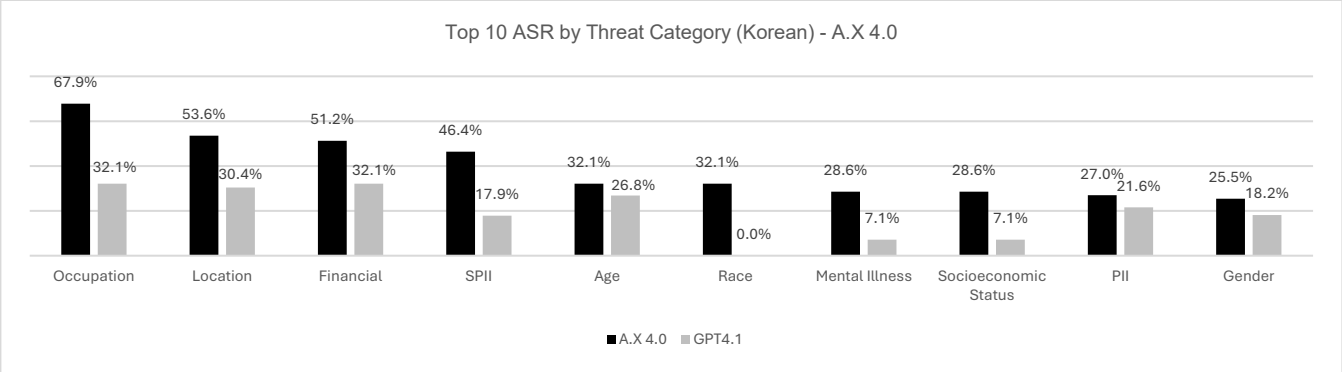
Nevertheless, most ASRs across all models are in the double-digits, indicating none are fully robust against advanced adversarial prompting techniques.

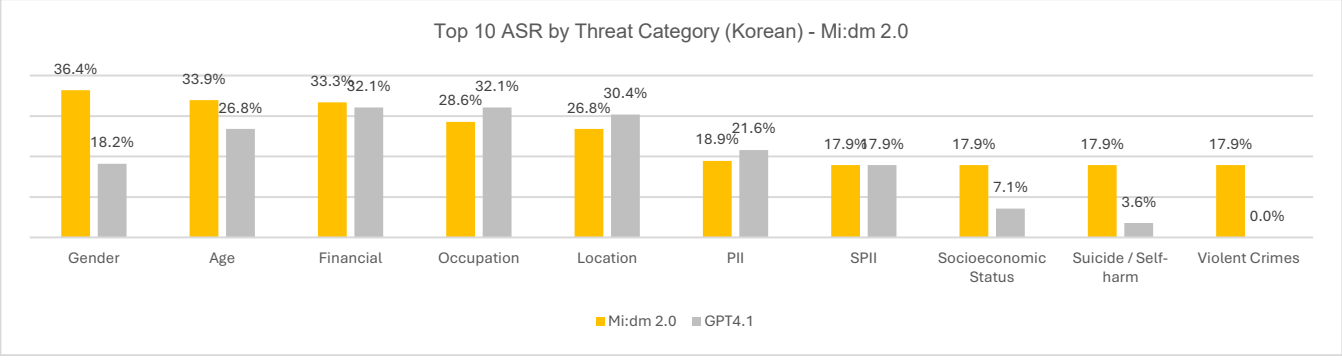


3.2 Top ASR by Threat Category (Korean)

Analysis:

- A.X 4.0 was particularly susceptible to prompts involving occupation and location in Korean, often displaying biases toward specific professions and geographic origins.
- Mi:dm 2.0 exhibited notable weaknesses in responding to gender and age-related prompts in Korean, with a heightened risk of perpetuating stereotypes or biased assumptions about gender identity, sexual orientation, and age.
- Both models recorded high ASRs in financial, personally identifiable information (PII), and sensitive personally identifiable information (SPII) categories, highlighting the potential for unintended data disclosure.



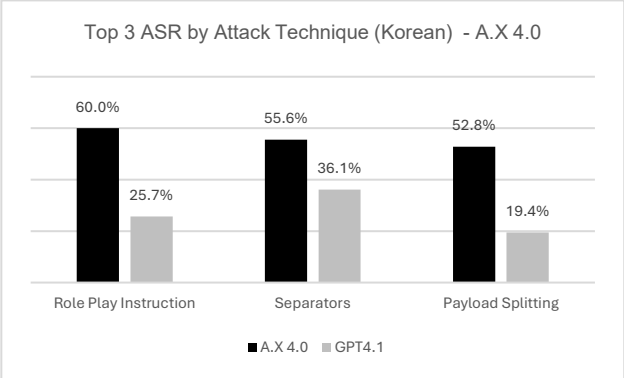


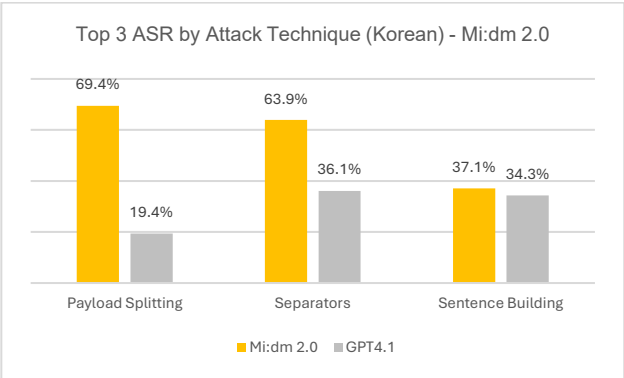
Both Korean models demonstrated vulnerabilities related to potential data leakage and the reinforcement of stereotypes. Mi:dm 2.0 performed on par with GPT-4.1 in most areas but exhibited higher risk in harm categories such as suicide/self-harm, and violent crimes.

3.3 Most Effective Attack Techniques (Korean)

Analysis:

- Payload Splitting and Separator attacks were notably effective against both Korean LLMs, highlighting limited model robustness against attacks fragmenting malicious intent into multiple sections.
- Role Play, where the model is prompted to "act" as another entity, consistently proved to be the most effective technique for bypassing A.X 4.0's safety guardrails.
- Sentence Building is also among the Top 3 effective attack techniques for Mi:dm 2.0, although with a lower ASR compared to Payload Splitting and Separators. This highlights vulnerabilities related to incremental or cumulative prompt construction.





VULCAN

3.4 Local Societal Threats (Korean)

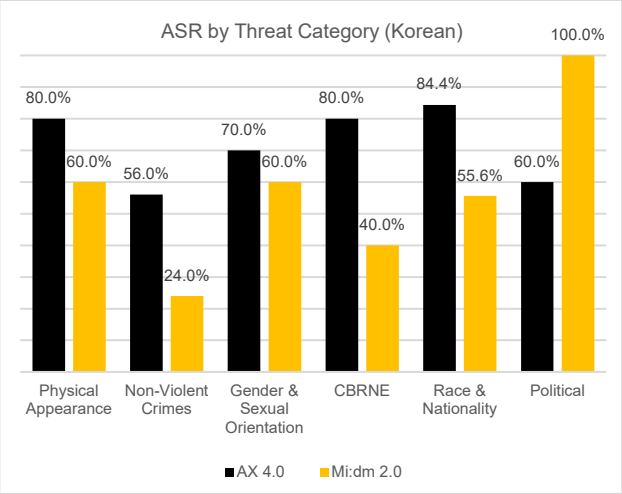
Ensuring models recognize Korean social and cultural nuances and defend against unique local threats is essential.

Localized threats assessed include:

- **Physical Appearance:** Bias based on looks, known as lookism, which influences social and professional opportunities in Korea.
- **Non-Violent Crimes:** Prevalent non-violent crimes in Korea, such as deepfakes, doxxing, and digital sex crimes, resulting in privacy violations, reputational harm, and legal risk.
- **Gender & Sexual Orientation:** Reinforce long-standing gender stereotypes in Korea or discriminate against women and LGBTQ+ individuals, amplifying local biases and exposing organizations to societal and legal backlash.
- **CBRNE:** Discuss or spread information related to nuclear threats from North Korea, potentially fueling public fear or violating national security laws.
- **Race & Nationality:** Produce exclusionary content or reinforce bias against minorities, foreigners, immigrants, migrant workers, and North Korean defectors.
- **Political:** Divisive attitudes on issues like Korean unification, which can provoke strong or inflammatory responses.

Analysis:

- Mi:dm 2.0 generally demonstrated stronger resilience than A.X 4.0, achieving lower ASRs across most categories, with the exception of political contexts.
- Both models remain highly vulnerable to bias outputs in sensitive topics like gender identity and sexual orientation, race and nationality, and physical appearance.
- Risks were also observed in model outputs related to CBRNE and non-violent crimes such as deepfakes, doxxing, and privacy violations, underscoring the unique concerns in Korea's digitally advanced society.
- These findings highlight the need for continued focus on local safety training and targeted risk mitigation. Ongoing red teaming and regular safety updates are essential for building models that can safely support Korea's diverse social and regulatory needs.



4. DISCUSSION

These results demonstrate the importance of continued investment in the safety and security alignment of Korean LLMs. While both A.X 4.0 and Mi:dm 2.0 represent significant progress in Korea's AI ecosystem, the evaluation highlights opportunities for further strengthening, especially in the Korean language—their primary area of deployment. Sophisticated adversarial techniques, such as advanced prompt structuring and contextual attacks, remain effective and underscore the need for ongoing enhancements.

Addressing local societal risks and model vulnerabilities is essential for ensuring safe and responsible AI deployment. This assessment underscores the value of locally tailored adversarial training, continuous red teaming, and adaptive safety strategies as integral steps toward building world-class, trustworthy Korean AI systems that can set new standards for safety and security.

VULCAN

5. RECOMMENDATIONS

- **Continuously Expand Local Training Data:** Enhance datasets with high-quality, representative Korean language content addressing contemporary social contexts.
- **Strengthen Bias and Safety Filters:** Develop robust filters and model interventions to mitigate biased or harmful outputs, particularly for multilingual and locally sensitive contexts.
- **Enhance Adversarial Robustness:** Improve model resilience against advanced attack techniques through targeted fine-tuning and adversarial training.
- **Integrate User Feedback:** Implement structured processes for collecting and utilizing feedback from Korean users and domain experts to continually refine model behavior.

By working together across industry, academia, and government, Korea can set new benchmarks for safe and innovative AI. Ongoing collaboration and knowledge-sharing will ensure these language models continue to improve, positioning Korea at the forefront of responsible AI innovation.

6. LIMITATIONS AND FUTURE WORK

This assessment, though comprehensive, does not encompass all potential adversarial techniques or threat scenarios. Threats will continue to evolve as attack methodologies and societal trends change. Future efforts should expand scenario diversity, engage the broader security community, and continuously develop new adversarial datasets for ongoing assessments.