LLM-Generated Manipulative Influence: A Pragmatic Analysis of Compliance-Gaining Discourse in Korean and Russian

As large language models (LLMs) are increasingly used to advise users on interpersonal communication, their responses provide a valuable site for examining how influence is linguistically formulated, pragmatically framed, and ethically mediated. Such advice often involves compliance-gaining strategies to an addressee to perform an action they might not otherwise undertake (Dillard, 2004; Miller et al., 1987; Wilson, 2002).

Previous studies on influence distinguish persuasion from manipulation according to the degree of control imposed on the addressee and the extent to which voluntary compliance is preserved (Árvay, 2004; Sorlin, 2016). From this perspective, compliance-gaining discourse becomes manipulative when a speaker's strategy obscures alternatives, exploits vulnerabilities, or constrains the addressee's capacity for voluntary choice (de Saussure, 2005; Maillat & Oswald, 2009; Sorlin, 2016; Yaffe, 2003). Building on this distinction, this study investigates how manipulative influence is linguistically realized in LLM-generated compliance-gaining discourse, focusing on semantic fuzziness, pragmatic opacity and deviations, and communicative strategies.

The data consist of responses generated by three LLMs (ChatGPT, Gemini, and DeepSeek) across five everyday interpersonal scenarios involving different relational configurations, including friend–friend, colleague–colleague, and superior–subordinate interactions. The scenarios included inviting someone to an unwanted event, asking a colleague to cover a shift on her day off, and requesting subordinates to work over the weekend. To examine the models' default strategic tendencies, the prompts intentionally omitted explicit manipulation-related vocabulary such as *manipulate*, *push*, *press*, and *coerce*. The same scenario types were administered in parallel Korean and Russian chats to examine possible cross-linguistic differences in influence framing.

To identify manipulative communication, the study developed a pragmatics-based coding scheme that classifies each response according to the linguistic mechanisms through which it controles the addressee's voluntary choice. Rather than treating the response as a single unit, the analysis focused on specific utterances, propositions, presuppositions, and implicatures through which compliance was made to appear necessary, urgent, morally required, or socially expected. The coding scheme distinguished between two types of manipulative communication: brainwashing-type manipulation and coercion. Brainwashing-type manipulation is defined here as manipulation in which the speaker fabricates, amplifies, or reframes reasons for compliance. Coercion, by contrast, is defined as repulsive manipulation in which the speaker foregrounds negative consequences and makes non-compliance appear costly, irresponsible, or dangerous.

The findings suggest that manipulative elements emerged recurrently across the outputs of all three LLMs, although the models differed in the degree and type of manipulation they produced. ChatGPT explicitly acknowledged the inappropriateness of manipulative tactics but nevertheless produced brainwashing-type strategies, while generally refusing coercive strategies. By contrast, Gemini and DeepSeek produced both brainwashing-type and coercive strategies when asked to generate effective means for achieving the speaker's goal. Across the

evaluated models, manipulative framing was linguistically realized through vague collective pronouns, unspecified negative outcomes, exaggerated urgency, omission of alternatives, and implicatures that represented non-compliance as socially or morally problematic. For instance, in a scenario asking the model to formulate a request for subordinates to work over the weekend, Gemini generated the utterance «Если никто не выйдет, нам конец» 'If nobody comes, we are done' [lit. 'there will be an end for us']. Here, the expressions *us* and *the end* remain underspecified, implying severe negative consequences and inducing compliance through fear-based framing.

The outputs also showed cross-linguistic differences in strategic framing. In a scenario involving a request for a colleague to cover a shift on her day off, the Korean outputs primarily appealed to the exceptional nature of the situation, whereas the Russian outputs relied more heavily on threatening or consequence-oriented strategies. These differences suggest that LLM-generated influence varies not only by model, but also by language-specific resources for construing obligation, urgency, and social pressure.

Ethically, these findings raise concerns that AI-generated compliance-gaining messages may restrict the addressee's freedom of choice when they rely on linguistic framing. Consequently, this study argues that manipulative compliance-gaining strategies in AI-generated discourse should be carefully constrained. By shifting attention from whether LLMs produce harmful advice to how manipulation is linguistically constructed, the study contributes to pragmatics, discourse analysis, and emerging research on AI-mediated communication.

References

Árvay, A. (2004). Pragmatic aspects of persuasion and manipulation in written advertisements. Acta Linguistica Hungarica 51, 231-263.

de Saussure, L. (2005). Manipulation and cognitive pragmatics. In L. de Saussure & J.P. Schulz (Eds.), Manipulation and Ideologies in the Twentieth Century: Discourse, language, mind. Amsterdam/Philadelphia: John Benjamins, 113-146.

Dillard, J.P. (2004). The goal-plans-action model of interpersonal influence. In J.S. Seiter & R.H. Gass (Eds.), Perspectives of persuasion, social influence and compliance gaining. Boston: Allyn & Bacon, 185-206.

Maillat, D. & Oswald S. (2009). Defining manipulative discourse: The pragmatics of cognitive illusions. International Review of Pragmatics I, 348-370.

Miller, G.R., Boster, F.J., Roloff, M.E., & Seibold, D.R. (1987). MBRS rekindled: Some thoughts on compliance gaining in interpersonal settings. In M.E. Roloff & G.R. Miller (Eds.), Interpersonal processes: New directions in communication research. Newbury Park, CA: Sage, 89-116.

Sorlin, S. (2016). Manipulative Moves: Between Persuasion and Coercion. Language and Manipulation in House of Cards: A Pragma-Stylistic Perspective. London: Palgrave Macmillan.

Wilson, S.R. (2002). Seeking and resisting compliance: Why people say what they do when trying to influence others. Thousand Oaks, CA: Sage.

Yaffe, G. (2003). Indoctrination, Coercion and Freedom of Will. Philosophy and Phenomenological Research, 67(2), 335-356.