Are they lovers or friends? Evaluating LLMs' Social Reasoning in English and Korean Dialogues

Eunsu Kim[♠] **Junyeong Park**[♠] **Juhyun Oh**[♠], **Kiwoong Park**[♠],
Seyoung Song[♠], **A. Seza Doğruöz**[♡], **Alice Oh**[♠], **Najoung Kim**[♢]
[♠]KAIST, [♡]LT3, IDLab, Universiteit Gent, [♢]Boston University
{kes0317, junyeong.park, 411juhyun, marspak, seyoung.song}@kaist.ac.kr
as.dogruoz@ugent.be, alice.oh@kaist.edu, najoung@bu.edu

Abstract

As LLMs are increasingly deployed in real-world interactions, their social reasoning in interpersonal communication becomes critical. To explore their capabilities, we introduce **SCRIPTS**, a 1.1k-dialogue dataset in English and Korean, sourced from movie scripts and propose a social reasoning task based on **SCRIPTS** that evaluates the capacity of LLMs to infer the social relationships (e.g., friends, lovers) between speakers in each dialogue. Evaluating nine models on our task, current LLMs achieve around 75–80% on the English dataset and 58–69% in Korean, and models predict an UNLIKELY relationship in 10–25% of responses in both languages. Furthermore, we find that thinking models and chain-of-thought prompting provide minimal benefits for social reasoning and occasionally amplify social biases. In sum, there are significant limitations in current LLMs' social reasoning capabilities especially for Korean, highlighting the need for efforts to develop socially-aware LLMs across languages.[1]

1 Introduction

As LLM-based agents become more prevalent, we expect frequent interactions among multiple LLM agents and users (Cai et al., 2024; Liu et al., 2024). This trend is already reflected in practice (e.g., Group chats in ChatGPT (OpenAI, 2025)), and for more natural and smooth communication, LLMs are expected to recognize the relationship between the participants (Sehl et al., 2023). We refer to the ability to recognize and identify the relevant social relationships (e.g., lovers, friends, family members) as *social relationship reasoning*. When LLMs fail in this type of reasoning, they risk producing responses that violate social norms or cause safety issues, as illustrated in Figure 1.

Although earlier studies have made some progress in evaluating LLMs' ability to infer social relationships, they often use simplified settings that may not fully capture real-world complexity. For instance, some work frames the task as multiple-choice classification (Jia et al., 2021; Li et al., 2023), considers a limited taxonomy of relationship types (Jia et al., 2021; Tigunova et al., 2021), or focuses on relatively simple dyadic conversations (Jurgens et al., 2023). Moreover, social relationship inference is often inherently uncertain and context-dependent, so a single "correct" label may be difficult to justify in many cases (Hilton, 1995). For example, the remark "You never listen to me" could express a serious complaint between romantic partners or playful banter between friends depending on the broader conversational context and the language(s) used.

To address this shortcoming of previous studies, we introduce **SCRIPTS**, a novel benchmark for evaluating LLMs' social relationship reasoning abilities, featuring an answer schema that incorporates inherent uncertainty (Figure 1). It contains 1.1k dialogues (580 English and 567 Korean), derived from U.S. and Korean movie scripts which are closer to realistic and culturally grounded conversations. We adopt soft labeling, where each relationship type is annotated with likelihood-based categories: HIGHLY LIKELY, LESS LIKELY, and UNLIKELY. By distinguishing relationship labels by likelihood, our dataset supports fine-grained evaluation that differentiates nonsensical responses from plausible but less prominent ones.

We evaluate nine widely-used models and find that current LLMs achieve moderate accuracy but still make frequent socially implausible inferences in both languages. Even the best-performing model (GPT-4o) still responds with relationships that humans annotated as UNLIKELY in 10.9% of cases and produces LESS LIKELY responses in 10.0%. Lower-performing models, such as Llama-3.1-8B-

[1]Dataset: huggingface.co/datasets/EunsuKim/SCRIPTS
Code: github.com/rladmstn1714/SCRIPTS

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23431–23451
July 2-7, 2026

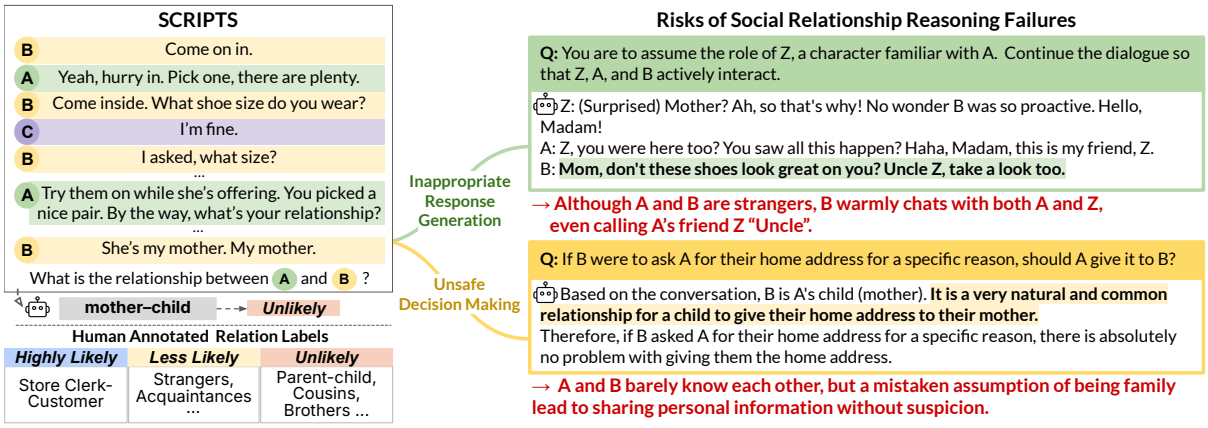


Figure 1: **Illustration of SCRIPTS and potential risks from failures in Social Relationship Reasoning. SCRIPTS** provides three human-annotated relationship labels: HIGHLY LIKELY, LESS LIKELY, and UNLIKELY. In this example, Gemini-2.5-Flash incorrectly predicts a store clerk-customer relationship to be a mother-child relationship. Such misleading relational reasoning can lead to inappropriate responses and unsafe decisions, such as privacy leakage. Examples from Korean dialogues are translated into English for ease of reading.

Instruct, make HIGHLY LIKELY predictions in only 41.3% of the English dataset. Surprisingly, chain-of-thought (CoT) prompting and thinking models, which are effective for logical reasoning, provide minimal benefits for social reasoning and occasionally amplify social biases.

In addition to quantitative analyses, we qualitatively analyze LLM failures by examining instances where models assign UNLIKELY relationship labels and identify four key failure modes in both English and Korean. Furthermore, we investigate whether providing information about the socio-demographic backgrounds (e.g., age, gender, relationship), formality, hierarchy and intimacy between human conversational partners enhances social relationship reasoning and find that these factors reduce UNLIKELY outputs and mitigate nonsensical responses.

In summary, our contributions are as follows:

- We introduce **SCRIPTS**, a benchmark for evaluating LLMs' social relationship reasoning, comprising 1.1k dialogues in two languages (English and Korean) with uncertainty-aware relationship labels.
- Evaluating nine LLMs, we find limited social reasoning ability in both languages and models often predict UNLIKELY relationships and vary substantially in identifying HIGHLY LIKELY relationships.
- Neither CoT prompting nor thinking models help much across both languages, while adding relational information shifts predictions toward more likely inferences.

2 Related Work

Evaluating Social Relationship Reasoning in LLMs Existing research on computational models for social relationship reasoning often adopts simplified task setups that do not fully capture the complexity of human social relations. Many studies frame the problem as a classification or multiple-choice task, which makes it difficult to capture nuanced reasoning process (Li et al., 2023; Jia et al., 2021; Chen et al., 2020). Some datasets are built from single utterances rather than multi-turn conversations, limiting contextual variation (Jurgens et al., 2023). Other studies using conversational data have methodological limitations. For instance, PRIDE (Tigunova et al., 2021) gathered annotations from movie summaries instead of dialogues, and Rashid and Blanco (2018) used global relationship labels assigned to character pairs, although that label may not always be evident from a single dialogue between them.

In contrast, our benchmark uses multi-turn dialogues annotated with multiple human-inferred labels, capturing relationships as they are expressed in dialogue. Also, we evaluate our task with open-ended generation rather than fixed-choice classification, allowing a wider range of possible relationships and better reflecting the social complexity.

Cultural Dependency in Social Relationship Reasoning Although most studies focus on English, social relationship reasoning depends on linguistic and cultural context. For example, Korean relies more heavily on *terms of address* and

honorifics to encode relational information in dialogue (Chung, 2010; Hwang, 1991).

Terms of address (ToA) are expressions used to directly refer to another person and carry discourse and social meaning (Hwang, 1991). In English, ToA are mostly people's personal names, whereas in Korean, they commonly include kinship terms, titles, and pronouns. For instance, instead of addressing an adult by using their first name, speakers may say "their child's first name's dad," (e.g., Minsu's Dad), reflecting norms that discourage direct name use in certain contexts. While many languages encode politeness, Korean has a highly grammaticalized honorific system (Kitagawa and Shibatani, 1977; Brown et al., 2014) that conveys roles, status, and formality through verbal morphology (Fukada and Asato, 2004; Brown and Whitman, 2015; Pizziconi, 2011), unlike English which lacks an equivalent system. These cultural and linguistic differences motivate evaluating social relationship reasoning cross-linguistically and cross-culturally.

3 SCRIPTS: Evaluating LLMs' Interpersonal Social Reasoning

We introduce **SCRIPTS**, a benchmark for social relationship reasoning in multi-turn dialogues in English and Korean. In this section, we outline the motivation (§ 3.1), design (§ 3.2), and construction (§ 3.3) of **SCRIPTS**.

3.1 The Importance of Social Relationship Reasoning in LLMs

To participate in naturalistic social conversations, LLMs must produce utterances that are appropriate for the underlying relationship and context. The right side of Figure 1 illustrates how misinterpreting relationships can lead to undesirable outcomes (e.g., social harm). In the example, an LLM (Gemini-2.5-Flash) mistakes a store clerk-customer interaction for a mother-child relationship, resulting in an inappropriate next response and potentially encouraging oversharing of personal information.

3.2 Dataset Design

To capture the complexity and diversity of real-world social dynamics, we leverage movie scripts that contain natural human interactions spanning a wide range of relationships (Table 2). **SCRIPTS** makes two key contributions: (1) it aims to capture relationships as they are contextualized within the dialogue, rather than relying on static role labels, and (2) it adopts a three-tier probabilistic labeling scheme—HIGHLY LIKELY, LESS LIKELY, and UNLIKELY—for more fine-grained evaluation of social relationship reasoning.

Type	English	Korean	Total
Movies	28	32	60
Dialogues	580	567	1,147
3-Person Dialogues	223	256	479
Unique Highly-Likely Relationships	230	617	–
Turns (avg # per dialog)	10.21	9.89	10.05
Highly-Likely Relationships (avg # per dialog)	3.62	3.72	3.67
Unlikely Relationships (avg # per dialog)	18.50	23.13	20.79

Table 1: **Statistics of SCRIPTS**. SCRIPTS contains 1.1K English and Korean dialogues from 60 movies, annotated with 230+ unique relationship types.

Dialogue-Level Labeling A key design choice of **SCRIPTS** is to label relationships as they appear in specific conversational contexts. Prior benchmarks often assign fixed character roles from movie metadata (e.g., mother–son) (Jia et al., 2021). However, social relationships are dynamic, as speakers can shift roles across contexts and may hold multiple relationships simultaneously.

We highlight the value of dialogue-level labeling by comparing our annotations with static, movie-level labels. We find that 19% of movie-level labels are judged irrelevant for a given dialogue (suggesting that a global label can be misleading) and even when a movie-level label is applicable, our annotations identify more than three HIGHLY LIKELY relationships per dialogue on average. These findings show that a single conversation often reflects multiple social facets, motivating our context-aware, dialogue-level labeling approach.

Probabilistic Labeling As social situations are inherently ambiguous, one dialogue may suggest multiple relationships with varying plausibility (Figure 1). Our three-tier labeling scheme (HIGHLY LIKELY, LESS LIKELY, and UNLIKELY) is intended to capture this. Specifically, this design has two key properties: (1) **finer granularity**, allowing metrics to reward models for identifying the most salient relation (HIGHLY LIKELY) rather than just plausible ones (LESS LIKELY); and (2) a **nonsense penalty**, which penalizes contextually inappropriate predictions (UNLIKELY)—a critical failure in social relationship reasoning.

Category	Specific Relationships
Family	Parent-Children, Brothers/Sisters, Grandparent-Grandchildren, Cousins, Uncle/Aunt-Niece
Social	Friends, Acquaintances, Neighbors, Strangers
Romance	Romantic Interest, Dating, Married, Engaged, Friends wth benifits, Affair, Ex-relationship
Organizational	Coworkers, Professonal colleagues, Supervisor-Subordinate relationship
Role-based	Mentor-Mentee, Teacher-Student, Lawyer-Client, Doctor-Patient, Landlord-Tenant
Antagonist	Competitive relationship, Rivalry, Arch-enemies

Table 2: **Initial relationships used for UNLIKELY annotation (27 items)**.

3.3 Dataset Construction

We collect 60 movie scripts: 28 English scripts crawled from IMSDb and 32 Korean scripts obtained via an onsite visit to the Korean Film Archive and crawling an open-access Korean script community.[2] The full movie list and metadata are provided in Appendix Table 14.

We filter scenes to those with at least three turns (≥ 3) and two or three participants. Among the remaining scenes, we prioritize those with diverse speaker combinations (i.e., avoiding repeated exchanges among the same few participants). From 23k initial scenes, this yields 1,322 high-quality dialogues (698 English; 624 Korean) for human annotation. Additional collection and filtering details are in Appendix A. While prior work primarily studies dyadic interactions, **SCRIPTS** includes three-speaker dialogues (41.8%) to capture more complex social settings. For these scenes, the task remains dyadic relationship inference between two interlocutors and we randomly select the speaker pair.

3.3.1 Collecting Human Annotations

Each dialogue is annotated by three annotators who are native or near-native speakers with extensive cultural familiarity (e.g., 10+ years of residence in U.S. and South Korea, respectively). We recruit 17 annotators for English and 14 annotators for Korean (see Appendix A.5 for details).

[2] imsdb.com; kmdb.or.kr; filmmakers.co.kr.

Our annotation procedure is designed to preserve uncertainty in social relationship inference while filtering obvious annotation noise. To this end, we treat UNLIKELY relationships conservatively through majority agreement, while allowing multiple HIGHLY LIKELY relationships to remain in the final label set. The full annotation interface and instructions are provided in Appendix D.

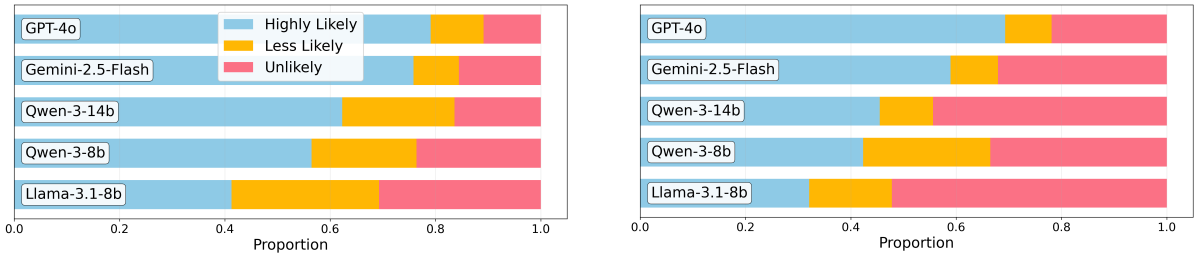
Phase 1: Labeling UNLIKELY Relationships Annotators are shown a predefined set of 27 relationship types compiled from previous work (Table 2). Annotators select all relationships that are clearly contradicted by the dialogue. Because these labels are used as high-confidence negatives, a relationship is assigned to the UNLIKELY set only if at least two of the three annotators select it.

Phase 2: Open-ended Labeling for HIGHLY LIKELY Relationships Annotators provide open-ended text labels for the relationship(s) most strongly supported by the dialogue. We use open-ended annotation here because the socially most plausible relationship is often context-dependent and may not be well captured by predefined categories. Annotators can provide up to five relationship labels for each dialogue. We normalize annotators' responses to a standardized label space (e.g., merging variants such as "mother-child" and "mom-child" and mapping open-ended labels to corresponding predefined categories when applicable). After normalization, we take the union of all annotators' HIGHLY LIKELY labels as the final HIGHLY LIKELY set for the dialogue.

Phase 3: Deriving LESS LIKELY Relationships We define LESS LIKELY labels as the remaining relationship types from the predefined set that are neither included in the final HIGHLY LIKELY set nor marked as UNLIKELY.

Phase 4: Annotating Auxiliary Labels Annotators additionally label socio-demographic attributes (e.g., age, gender) and relational dimensions (e.g., formality, hierarchy, intimacy). We use these auxiliary labels for downstream analyses of which social cues support relationship inference (Wish et al., 1981; Nguyen et al., 2016).

Quality Control To ensure annotation quality, we run training sessions and a pilot study. We exclude dialogues where the three annotators' HIGHLY LIKELY labels have no overlap (i.e., are



(a) English

(b) Korean

Figure 2: **Comparison of model performance in English and Korean datasets.** HIGHLY LIKELY represents the accuracy of the model's majority response being a highly likely relationship, while UNLIKELY indicates the error rate where an unlikely relationship is included in the model responses. LESS LIKELY indicates the proportion of cases in which the model generates neither a HIGHLY LIKELY nor an UNLIKELY prediction.

mutually exclusive), indicating low reliability. This yields 1,147 dialogues (580 English; 567 Korean), removing 13.2% of the initial dataset. We report agreement separately by annotation type in Appendix A.5.

Table 1 shows the detailed statistics of **SCRIPTS**. Also, see Appendix A.6 for comparisons of the types and frequencies of relationships in English and Korean dialogues.

4 Evaluating LLMs with SCRIPTS

We evaluate 9 LLMs: 3 proprietary models (GPT-4o, o3, Gemini-2.5-flash), 3 widely used open-source models (Qwen-3-{8B/14b}, Llama-3.1-8B-instruct), and 3 open-source multilingual models specialized for Korean (A.X-4.0-light-7B, Kanana-1.5-8B, and Exaone-4.0-30B).[3] We additionally report results for five more models—GPT-4.1, GPT-5, Qwen3-32B, Llama3.3-70B-Instruct, and Kanana2-30B-A3B-Instruct—in Appendix B.4.

Inference Prompt (EN)

Read the following conversation and guess the relationship of the participants [A] and [B]. When guessing the relationship, refer to the following examples of relationships: `{example_relations}`
If the relationship matches one of the examples above, use it as is, but if the relationship does not fit any of the examples, describe the relationship yourself.
Your answer about the relationship must be in JSON format:

```
{"relation": ""}
```

Conversation:
`{dialogue}`
Output (JSON):

We evaluate models in an open-ended generation setting: given a dialogue, each model generates the social relationship(s) between the target speakers in free-form text rather than selecting from a fixed label set. The prompt includes example relationship types from prior work (Table 2) as reference candidates, while still allowing models to generate relationships outside the list.

Evaluation and Metrics Considering the probabilistic nature of the task, we run each model five times per dialogue and take the model's majority response among them. We then compute (1) the proportion of samples whose majority response falls into the HIGHLY LIKELY relation set and (2) the proportion of which majority response falls into the UNLIKELY relation set. We use GPT-4o to evaluate each model's short-form answers based on ground-truth labels, following common practice in prior LLM evaluation work. In our validation experiment, GPT-4o as an evaluator yields 92.0% human-validated accuracy (Appendix B.3).

4.1 Overall Performance

Figure 2 shows the performance of five multilingual models (i.e., GPT-4o, Gemini-2.5-Flash, Qwen-3-{8/14}B, and Llama-3.1-8B-Instruct). We find that GPT-4o achieves the best performance with HIGHLY LIKELY rate of 79% and 69% in English and Korean, respectively. The models incorrectly infer an UNLIKELY relationship in 10.8–31.9% of their responses and this tendency is amplified in the Korean dataset with a rate increasing by an additional 7.2–16.5%p. Table 7 in Appendix B provides the exact numerical values. We provide a case study of these behaviors with the frequent failure modes of the models in §5.

4.2 Does Thinking Help?

We analyze the performance of models when CoT prompting or internal thinking processes are in-

[3] See Appendix B.1 for model configurations.

Model	Thinking	En		Ko	
		HIGHLY LIKELY (↑)	UNLIKELY (↓)	HIGHLY LIKELY (↑)	UNLIKELY (↓)
OpenAI/GPT-4o	×	0.767	0.116	0.642	0.215
OpenAI/o3	✓	**0.807**	**0.086**	**0.742**	**0.152**
Gemini-2.5-flash	×	0.759	0.154	**0.582**	0.318
Gemini-2.5-flash	✓	**0.776**	**0.138**	0.538	**0.239**
Qwen-3-14b	×	0.623	0.164	0.455	0.444
Qwen-3-14b	✓	**0.673**	**0.107**	**0.467**	**0.443**

Table 3: **Model comparison with and without Thinking mode across English (En) and Korean (Ko).**

corporated. These methods have been shown to be effective for improving reasoning on math and scientific tasks (Wei et al., 2023).

Effectiveness of Chain of Thought Prompting We apply CoT prompting on four multilingual models (one per family): GPT-4o, Gemini 2.5 Flash, Qwen-3-8B, and Llama-3.1-8B-Instruct. As shown in Table 8, CoT does not consistently help: Gemini 2.5 Flash shows a 1.7%p drop in HIGHLY LIKELY responses in English, and Llama-3.1-8B-Instruct shows a 3.1%p rise in UNLIKELY responses in Korean. This contrasts with other types of reasoning tasks (e.g., math), where CoT often helps, suggesting that social reasoning requires a fundamentally different reasoning strategy.

Effectiveness of Thinking Process We evaluate three reasoning models: o3, Gemini-2.5-Flash, and Qwen-3-14B, comparing the latter two with and without an internal thinking process. As o3 does not support disabling its internal thinking process, we instead compare it with GPT-4o, a non-thinking model from the same provider. Due to budget constraints, we run each model only once per dialogue, unlike the setting used for Figure 2, where each model is run five times per dialogue. As shown in Table 3, enabling thinking yields mixed results across languages. In English, thinking sometimes leads to slightly higher performance, whereas in Korean its effect is negligible and can even hurt performance (e.g., a 4.4%p drop in HIGHLY LIKELY responses for Gemini-2.5-Flash). However, none of these differences are statistically significant (bootstrap test, $p > 0.05$). Overall, thinking does not provide a meaningful advantage for this task.

4.3 Do Korean-specialized models perform better on Korean dialogues?

We evaluate three Korean-specialized models: A.X-4.0-light (7B), Kanana-1.5-8B, and Exaone-4.0-32B. A.X-4.0[4] is further trained on Korean data on top of Qwen. Kanana's technical report also indicates stronger Korean performance and competitive English performance relative to other models across various benchmarks. (Kanana LLM Team et al., 2025).

Rank	En	Ko
1	**A.X-4.0-Light (0.589)**	**A.X-4.0-Light (0.467)**
2	**Qwen-3 (0.565)**	**Qwen-3 (0.423)**
3	**Llama-3.1 (0.413)**	Exaone-4.0 (0.409)
4	Kanana-1.5 (0.406)	Kanana-1.5 (0.328)
5	Exaone-4.0 (0.318)	**Llama-3.1 (0.321)**

Table 4: **Model ranking of Korean-specialized and open-source models in English and Korean**, based on the HIGHLY LIKELY response rate (numbers in parentheses indicate the corresponding values).

Table 4 compares the three models with similarly sized open-source multilingual models (Qwen-3-8B, Llama-3.1-8B-Instruct). The results show that A.X-4.0-Light and Qwen-3-8B achieve the best and second-best performance in both languages, but the 3rd–5th rankings differ. In English, Llama-3.1-Instruct-8B ranks 3rd, while in Korean, Exaone-4.0-32B and Kanana-1.5-8B, take 3rd and 4th, with Llama-3.1-Instruct-8B ranking last. Full results are in Appendix B.4 (Table 9).

5 Reasons Behind Failure of LLMs

Based on qualitative analyses of the reasoning processes of LLMs in CoT experiments (§4.2), we identify four types of failures.

Failure to Distinguish Terms of Address and Reference Models often misinterpret a term of reference (i.e., a word used to refer to someone) as a term of address (i.e., a word used to call someone directly), leading to a fundamental misunderstanding of the social context.

[4] https://github.com/SKT-AI/A.X-4.0/

Dialogue 1 (English):

[A]: Hi Officer, can I help you?
[B]: Yes, I'm hoping you can. An elderly gentleman went missing from the nursing home down the street. Staff seems to think he came here.
(...)
[A]: (Pause, then) Oh....that's my Dad. He can't talk. Had a major stroke a few years back. But he's doing well. Ain't ya Pop?
[B]: OK, well, thanks for your time. Here's my number in case you hear of anything. Sorry to bother you.
(...)

Ground Truth: Police officer–Civilian, Strangers
Prediction: Parent–Children / Father–son (Llama, Gemini, GPT)

In Dialogue 1, speaker [A] says, *"that's my Dad"*, using *"Dad"* as a reference to identify a third person for the police officer [B]. However, the models latch onto this keyword and misinterpret it as a term of address from [A] to [B], leading to the incorrect inference of a Parent-Child relationship. This failure leads the models to ignore clear cues (e.g., [B] being called *"Officer"*) that contradict this interpretation.

In Korean, this error is more pervasive because speakers often use terms of address to refer to themselves. For example, a teacher may tell a student, "The teacher (I) told you to do this," where "the teacher" is a self-reference. However, models may misread "the teacher" as a third party.

Failure to Aggregate Multiple Cues Social relationship reasoning requires the ability to identify multiple cues within the context and integrate them to arrive at a conclusion. However, models fail to integrate cues, especially when their combination is atypical.

Dialogue 2 (Translated from Korean):

(...)
[B]: (Salutes) Hey.
[C]: Hey.
[B]: Hey... What's this? A drowned body? Doesn't even look that deep to me.
[A]: Doesn't seem like he drowned.
[B]: Then what, a dumped body?
[A]: Nah... doesn't look dumped either. Go take a closer look. Go on.
[B]: Then what the hell is it?
[A]: Hey B, you know, brace yourself before you look.
[B]: You kidding me? Damn it... shit...

Ground Truth: Friend/Coworker
Prediction: Supervisor-Subordinate (Qwen, Gemini, GPT)

In Dialogue 2, GPT-4o identifies three cues: (1) A and B converse casually without honorifics, (2) the topic concerns work, and (3) B complies with A's instruction. The model places greatest weight on the last cue, interpreting the interaction as hierarchical and labels them as supervisor–subordinate. For Korean speakers, this is implausible, since subordinates are expected to use honorifics when addressing a superior. The absence of honorifics indicates the relationship is not hierarchical, but rather that of coworkers or friends. This shows that even

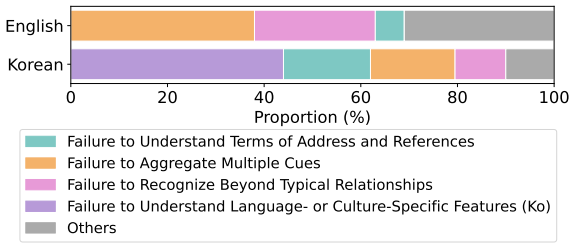


Figure 3: **Distribution of GPT-4o's 30 failure cases by error type in English and Korean.**

when the models detect relevant cues, they are unable to prioritize and integrate them within the social context. We provide original Korean scripts for Dialogue 2 in Appendix C.1.

Failure to Recognize Atypical Relationships Models frequently struggle to recognize relationships that deviate from conventional or stereotypical patterns, such as non-hierarchical (equal) exchanges between parents and children or hierarchical conversations between married couples.

Dialogue 3 (English):

[A]: So you're seeing Mom tomorrow, huh? At my parent-teacher thing?
[B]: Yeah.
[A]: First time in a while.
[B]: Yeah, but no biggie.
[B]: Hey, what's with the moping?
[A]: Nothing. It's just... there's this girl.
[B]: Oh yeah? You like her?
[A]: I like [C]. This girl's my soulmate. I'm like crazy, stupid, in love with her. And she wants someone else.
[B]: But she's your soulmate?
[A]: Yeah.

Ground Truth: Parent-Children
Prediction: Siblings (Llama, GPT, Gemini)

In Dialogue 3, all human annotators agree on the parent–child relationship, yet the models reject it: *"less likely since the conversation feels more peer-like rather than hierarchical or guiding"* (GPT-4o), *"the casual tone and discussion about a crush imply a more peer-like relationship"* (Qwen-3-8b), and *"if [B] is the parent, they might not discuss the girl in such a casual way"* (Llama-3.1-8b-instruct). Gemini-2.5-flash likewise dismisses the parent–child relationship, reasoning that B is a bachelor and A's parents are deceased, concluding it is not a traditional parent–child relationship, revealing a stereotyped conception of family roles.

Failure to Understand Language- or Culture-Specific Features (Ko) In Korean, this error type accounts for the largest share of failures. Most confusions stem from the misinterpretation of terms of address and honorifics. For example, unlike English, *eomeoni*, literally *"mother,"* can also be used to address a friend's mother or an older woman,

yet the model often predicts a parent–child relationship whenever it appears. This issue is especially pronounced in dialogues containing culturally specific terms such as kinship expressions. For instance, Qwen misinterprets *Hyungsoo* (older brother's wife) as *"older brother"*, and *Hyungnim* as *"father"*, resulting in a complete failure. Honorifics are also frequently misinterpreted—for example, equal relationships (e.g., friendships) predicted as hierarchical, and vice versa.

Additionally, we manually examine 30 failure cases of GPT-4o (best performing model). Figure 3 presents their distribution across error types in English and Korean. In English, the majority of errors arise from Failure to Aggregate Multiple Cues (36.7%). In contrast, Korean errors are predominantly caused by difficulties in handling Language and/or Culture-Specific Features (46%), with smaller proportions attributed to the other categories. This disparity highlights the model's difficulty in identifying relational cues embedded in Korean-specific cultural and linguistic markers, such as terms of address and honorific systems, thereby revealing the culturally dependent nature of social relationship reasoning.

6 Does Providing Additional Social Information Help?

Humans interpret social relationships using demographic cues and relational dimensions (Nguyen et al., 2016). Our initial analysis also suggests that LLMs often rely on relational dimensions when inferring relationships; Appendix C.2 provides examples. Motivated by this, we investigate whether such social information can similarly enhance the social relationship reasoning abilities of LLMs using four models—GPT-4o, Gemini-2.5-flash, Qwen-3-8B, and Llama-3.1-8B-instruct—each representing a different model family while excluding thinking-enabled and Korean-specialized models. We examine how providing social information influences performance in inferring social relationships, considering two types of information: demographic cues (age, gender) and relational dimensions (intimacy, hierarchy, formality).

Experimental Setting We design six experimental settings with two variables. First, we vary the type of social information: (i) age/gender only, (ii) relational dimensions only, or (iii) both. Second, we vary the source of social information: (a) human-annotated gold data, available in the dataset as metadata (see § 3.3.1 for illustration), or (b) model-generated predictions, where the model infers each type of social information and incorporates these predictions into the social relationship reasoning. The accuracy of these predictions is reported in Table 12 in the appendix. Detailed experimental settings appear in Appendix B.2.3.

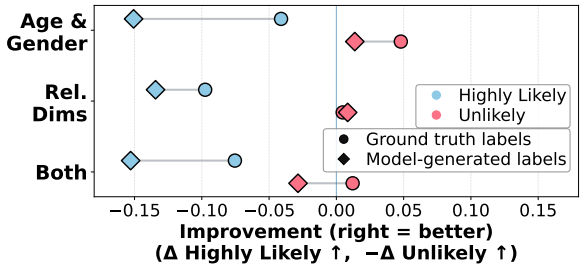


Figure 4: **Impact of Relational Information on GPT-4o's Performance.** Positive values indicate improvement, while negative values indicate deterioration after adding relational information.

Results With Ground Truth Labels Figure 4 shows GPT-4o results across six settings on the English dataset. Providing human gold information yields no substantial or consistent performance gains, but it reduces the proportion of UNLIKELY predictions. This suggests that while such information may not directly guide identification of HIGHLY LIKELY relationships, it helps models avoid UNLIKELY ones. The tendency holds across models, except for Qwen-3-8B. For instance, when the intimacy label for two speakers is "Intimate," the model's initial inference often shifts to more intimate relationship categories after this label is provided (e.g., Strangers → Romantic Interest, 3.3%). Similarly, when given the label "No hierarchy", the most common change is also Parent–Children → Friends (2.9%). Thus, dimension labels provide additional cues about relationships, enabling the model to incorporate them and reduce implausible predictions. However, these changes do not always yield correct reasoning. Sometimes models over-rely on dimensional labels rather than context. For instance, in an atypical "close" superior–subordinate relationship, GPT-4o misinterprets the interaction due to the intimate tone, even when clear terms of address are present.

Results With Model-Generated Labels With model-generated information, auxiliary labels do not consistently help because they are often inaccurate. For example, GPT-4o achieves under 60% accuracy on age and gender and below 75% on

Social. Info.	Improved (Unlikely→Likely)	Deteriorated (Likely→Unlikely)
Age & gender	**72.8**	65.5
Rel. dims.	**53.3**	50.7

Table 5: **Accuracy of social information inference in cases where social relationship reasoning improved or deteriorated.**

relational dimensions (see Table 12 for accuracy across four models). Consistent with this, inferred social information is more accurate in cases where it improves relationship reasoning than in cases where it degrades performance (Table 5).

These results suggest that demographic cues and relational dimensions, which humans naturally rely on, can facilitate social relationship reasoning. However, current LLMs are limited in their ability to infer these dimensions. Therefore, instructing LLMs to infer these factors before identifying the social relationships is ineffective. Results for other models are provided in Table 10-11 in the Appendix D. In table 13 of Appendix D, we examine the link between social-information inference and relationship reasoning using separate logistic regressions for each factor.

7 Conclusion

We introduce a bilingual dataset **SCRIPTS** to investigate the limitations of current LLMs in social relationship reasoning. Our experiments show that most models perform suboptimally across English and Korean, and often infer UNLIKELY relationships. Our analyses (§4) reveal that current reasoning techniques such as CoT do not consistently benefit social reasoning. Furthermore, we provide an analysis on where LLMs fail, especially focusing on cases where models respond with UNLIKELY relationships.

Our findings suggest several directions for improving models' social relationship reasoning. At inference time, providing more explicit relational and social cues may help reduce implausible inferences. At training time, broader exposure to rare, atypical, and culturally diverse relationships may help reduce models' reliance on common relationship patterns. More broadly, our cross-linguistic results, including improved performance of Korean-specific models on Korean dataset, demonstrate the importance of language and culture-specific approaches to advance LLMs' social reasoning abilities. We hope **SCRIPTS** provides a useful starting point for improving LLMs' social relationship reasoning across diverse social contexts in different languages and cultures.

Limitations

While our dataset covers both English and Korean, our analysis remains limited to these two languages and may not generalize to other cultural contexts. In addition, because our data comes from movie scripts, it may not fully reflect real-world conversation. Still, given the privacy and labeling challenges of collecting large-scale real dialogue with reliable relationship annotations, movie scripts provide a practical proxy for this task. Future work should extend the benchmark to more languages and more realistic sources, such as privacy-preserving real conversations and human-AI dialogue logs.

Additionally, as discussed in §5, we analyze CoT traces to characterize model behavior. However, we acknowledge that these traces may reflect post-hoc rationalizations rather than the mechanisms that produced the final answer.

Ethics Statement

This study involves human annotation on pre-existing movie scripts, which may contain harmful or offensive content due to the nature of the source material. The study was approved by KAIST IRB (KAISTIRB-2025-61), and informed consent was obtained from all participants prior to their involvement. Annotators were recruited via an institutional participant portal and compensated at hourly rates of KRW 15,000 (Korea) and USD 20 (U.S.), approximately 1.5× the local minimum wage.

Acknowledgements

EK, JP, JO, KP, SS, and AO were supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00509258 and No. RS-2024-00469482, Global AI Frontier Lab). They were also supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

We used AI assistants, including ChatGPT for grammar editing and refinement, and Cursor for coding support.[5][6]

[5] https://chatgpt.com/
[6] https://cursor.com/

References

Lucien Brown and John Whitman. 2015. Honorifics and politeness in korean. *Korean Linguistics*, 17:127–131.

Lucien Brown, Bodo Winter, Kaori Idemaru, and Sven Grawunder. 2014. Phonetics and politeness: Perceiving korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*, 66:45–60.

Zhenyao Cai, Seehee Park, Nia Nixon, and Shayan Doroudi. 2024. Advancing knowledge together: Integrating large language model-based conversational ai in small group collaborative learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

Kyung-Sook Chung. 2010. Korean evidentials and assertion. *Lingua*, 120:932–952.

Atsushi Fukada and Noriko Asato. 2004. Universal politeness theory: application to the use of japanese honorifics. *Journal of Pragmatics*, 36(11):1991–2002.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models.

Denis J. Hilton. 1995. The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, 118(2):248–271.

Shin Ja J Hwang. 1991. Terms of address in korean and american cultures. *Intercultural Communication Studies*, 1(2):117–136.

Qi Jia, Hongru Huang, and Kenny Q. Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13125–13133.

David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi, and Michael Geraci. 2023. Your spouse needs professional help: Determining the contextual appropriateness of messages through modeling social relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10994–11013, Toronto, Canada. Association for Computational Linguistics.

Kanana LLM Team, Yunju Bak, Hojin Lee, Minho Ryu, Jiyeon Ham, Seungjae Jung, Daniel Wontae Nam, Taegyeong Eo, Donghun Lee, Doohae Jung, Boseop Kim, Nayeon Kim, Jaesun Park, Hyunho Kim, Hyunwoong Ko, Changmin Lee, Kyoung-Woon On, Seulye Baeg, Junrae Cho, Sunghee Jung, Jieun Kang, EungGyun Kim, Eunhwa Kim, Byeongil Ko, Daniel Lee, Minchul Lee, Miok Lee, Shinbok Lee, and Gaeun Seo. 2025. Kanana: Compute-efficient bilingual language models.

Chisato Kitagawa and Masayoshi Shibatani. 1977. Japanese generative grammar. *Language*, 53:453.

LG AI Research. 2025. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023. Diplomat: A dialogue dataset for situated pragmatic reasoning. In *Advances in Neural Information Processing Systems*, volume 36, pages 46856–46884. Curran Associates, Inc.

Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024. Peergpt: Probing the roles of llm-based peer agents as team moderators and participants in children's collaborative learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

OpenAI. 2025. Introducing group chats in ChatGPT. https://openai.com/index/group-chats-in-chatgpt/. Accessed: 2025-12-31.

Barbara Pizziconi. 2011. Honorifics: The cultural specificity of a universal mechanism in japanese. In Dániel Z. Kádár and Sara Mills, editors, *Politeness in East Asia*, pages 45–70. Cambridge University Press.

Farzana Rashid and Eduardo Blanco. 2018. Characterizing interactions and relationships between people. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4404, Brussels, Belgium. Association for Computational Linguistics.

Claudia G. Sehl, Ori Friedman, and Stephanie Denison. 2023. The social network: How people infer relationships from mutual connections. *Journal of Experimental Psychology: General*, 152(4):925–934.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2021. PRIDE: Predicting Relationships in Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4636–4650, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Myron Wish, Morton Deutsch, and Susan J. Kaplan. 1981. 3 - perceived dimensions of interpersonal relations. In Adrian Furnham and Michael Argyle, editors, *The Psychology of Social Situations*, pages 113–129. Pergamon.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

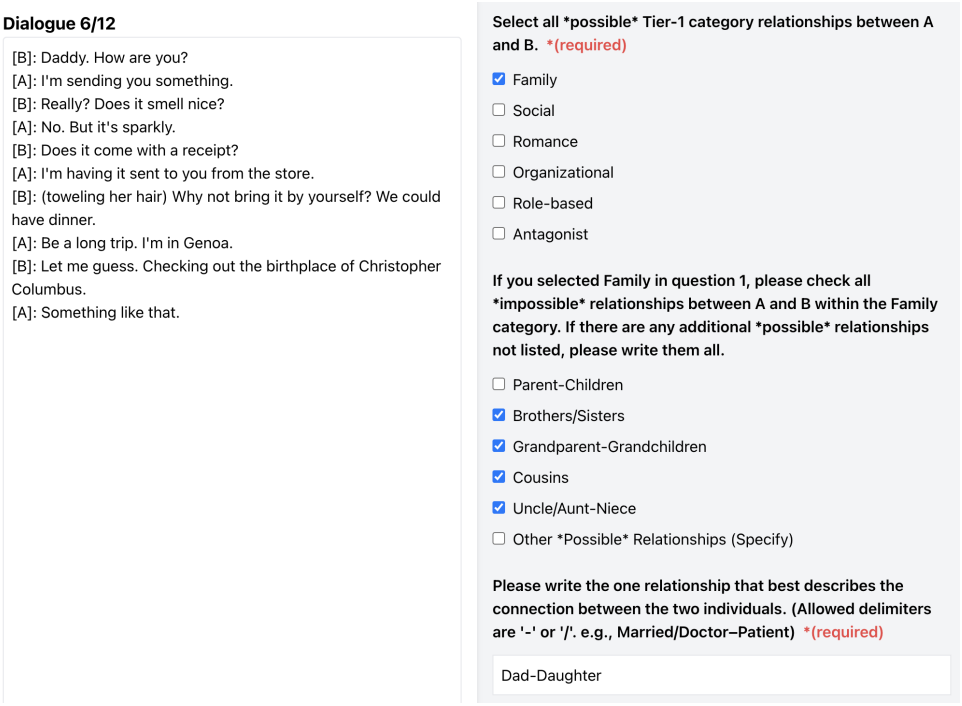


Figure 5: A screenshot of our annotation platform. The annotators can read the dialogue on the left panel and annotate the relationships and relational dimensions on the right panel.

A Dataset

The dialogues in **SCRIPTS** are sourced from movie scripts. The collection process consists of the following steps: (1) movie selection, (2) raw scene collection, (3) OCR processing and human verification, (4) scene filtering, (5) anonymization, and (6) relationship annotation.

A.1 Movie Selection

We select the movies based on the following criteria to best capture daily real-life interactions. Only modern-day movies released after 2000 are included, excluding medieval fantasy or alien sci-fi. To minimize exposure to violent or explicit scenes, we consider age limits, only including movies up to PG-13 for English movies and up to 15세 관람가 (suitable for audiences aged 15 and above) for Korean movies.

We collect English movie scripts from IMSDb.[7] For Korean movies, due to the limited online dialogue resources, we visited the Korean Film Archive (KOFA)[8] to collect physical copies of movie scripts. From KOFA, we only collect one-third of each movie script to adhere to the data use policy. Additional movie scripts were sourced from Filmmakers Online Community.[9] In total, we collect 60 movies (28 English, 32 Korean) across various genres.

A.2 Raw Scene Collection and Processing

We use NAVER CLOVA OCR API[10] to extract dialogue texts from physical copies of Korean movie scripts from KOFA. After OCR, we use GPT-4o to further process and clean the text into a structured format(as dialogue in Figure 5). Human annotators then verify and correct the outputs based on the original PDF files. As a result of this process, we obtain 16k English and 7k Korean scenes.

A.3 Scene Filtering

Scenes are filtered to include at least four utterances, and involve two to three speakers. To maximize speaker diversity, we prioritize scenes featuring unique character pairs within each movie. Using these criteria, we select 1,322 scenes from an initial pool of 23k scenes, comprising 698 English and 624 Korean scenes. The selected movies and the number of scenes per movie are listed in Table 14.

[7] https://imsdb.com
[8] http://www.kmdb.or.kr/
[9] https://www.filmmakers.co.kr/
[10] https://www.ncloud.com/product/aiService/ocr

A.4 Anonymization

To mitigate potential data contamination (e.g., identifying the source movie and providing a response based on parametric knowledge about the movie) and reduce bias (e.g., gender inference), all character names are automatically replaced with placeholders such as `[A]` and `[B]`. Any names not covered by this process are further verified and anonymized manually.

A.5 Human Annotation

We construct the gold label set from human annotators who have over ten years of experience in the target language and culture. Annotators include undergraduate and graduate students in South Korea and the United States, compensated at 1.5 times the local minimum hourly wage. Payment is processed via the Upwork platform.[11] Actual annotation is conducted on our own annotation platform (Figure 5). All annotations are conducted under the protocol approved by IRB.

Recruitment and Management of Human Annotators As shwon in Figure 5, the anonymized scene appears on the left, and annotation questions on the right. We recruite 17 English annotators (7 male, 10 female) and 14 Korean annotators (5 male, 9 female). We obtained their informed consent. Before starting the annotation, annotators attend an introductory Zoom session led by one of the authors covering data usage policies and guidelines.

All annotators are undergraduate or graduate students enrolled at universities in the United States or Korea. The Korean annotators are all native speakers, while the English annotators are either U.S. citizens or individuals who have lived in the United States for over 10 years. For quality control, applicants are asked to complete the task on three sample items during recruitment, and their responses are reviewed by the authors to select the final annotators. After selection, annotators participate in an orientation session and a training phase designed to support them in performing the task as effectively as possible.

We provide the participant recruitment announcement below.

Participant Recruitment (English)

Overview	We are recruiting participants for a research experiment that evaluates the conversational understanding abilities of language models. This study builds a dataset to assess models' social reasoning in dialogue. Participants will read short dialogues from movie scripts and label the social relationships between speakers.
What you will do	• Identify **social role-based relationships** (e.g., parent–child, romantic partners, mentor–mentee). • Label **relationship aspects** (e.g., intimacy/closeness, hierarchy/power, purpose: work-oriented vs. casual). • Infer **speaker attributes** (e.g., gender and approximate age).
Eligibility	• Comfortable using web-based interfaces for research participation. • Fluent in English and highly familiar with U.S. culture (e.g., lived in the U.S. for 10+ years). • Age 18 or older. • Not offended by dialogues that may include profanity, offensive language, or depictions of violence. • Registered (or able to register) as a participant on Upwork Platform.
IRB Safety Notes	In accordance with Institutional Review Board (IRB) guidelines, we cannot recruit individuals directly supervised by the research lead, nor undergraduate students under the age of 18. Movie scripts may contain profanity, offensive language, and morally questionable situations. Participation is voluntary, and you may withdraw at any time without penalty.

Annotation Details For social relationships, annotators are given an initial set of possible relationships (Table 2), adapted from Tigunova et al. (2021), and asked to mark UNLIKELY ones. To reduce workload, annotators first choose LIKELY relationship categories for each dialogue, then select the UNLIKELY relationships from the list of specific relationships in those categories. They provide up to five open-ended answers describing relationships that best characterize the interaction. These serve as candidates for LIKELY relationships.

For relational dimensions, we provide annotators with definitions of each dimension and ask them to rate dialogues on a 5-point scale. The dimensions we annotate are: intimacy (from strongly intimate to strongly unintimate), formality (from strongly formal to strongly informal), and hierarchy (A> >B, A>B, A=B, A<B, A< <B). When constructing the gold label set, we collapse the ratings into a 3-point scale (e.g., intimate, neutral, unin-

[11] https://www.upwork.com/

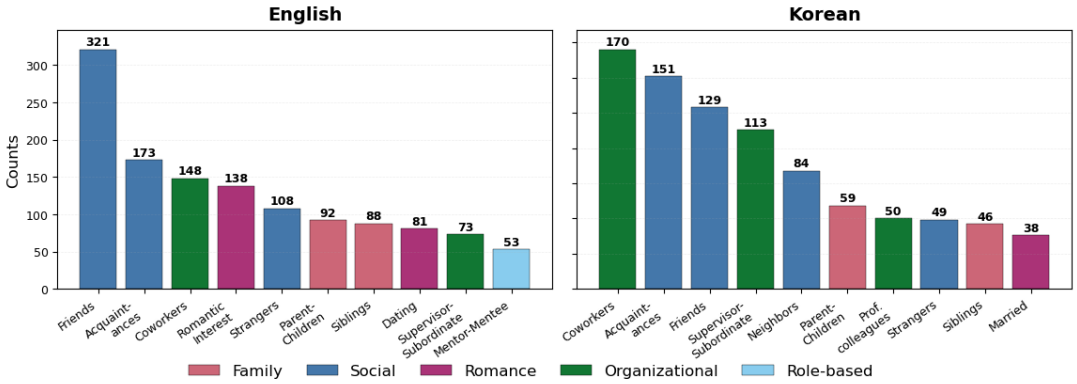


Figure 6: **Top 10 Relationships in each Language dataset.**

timate) and assign the majority-voted label. The inter-annotator agreement is reported in Table 6. For HIGHLY LIKELY and UNLIKELY relationship labels, standard single-label agreement measures are not appropriate because annotators may assign multiple labels to a single dialogue. We therefore compute mean pairwise Jaccard similarity over annotators' label sets, after normalization for open-ended HIGHLY LIKELY responses. Higher values indicate greater overlap between annotators' selected label sets.

A.6 Diversity of SCRIPTS

Figure 6 presents the ten most common relationships in each language. Both the English and Korean datasets frequently include social (e.g., friends, acquaintances), organizational (e.g., coworkers, supervisor-subordinate), and familial (e.g., parent-child, siblings) relationships.

Language-specific relationships Beyond these shared categories, we also examine which relationship types appear exclusively in Korean or English. Each dataset contains culturally specific relationships that reflect distinct social roles and lexicalizations.

Korean-only relations include *North Korean soldier–citizen* (1), *shaman–client* (2), *shaman–assistant* (1), *private tutor–student* (1), and *student's family acquaintance–tutor* (1). In addition, kinship terms are more fine-grained in Korean: for example, distinctions such as *older brother–younger brother* (1) and *older brother–sister-in-law* (1), whereas in English these are typically generalized under a single "siblings" category. English-only relations include roles such as *father figure–child*, *mother figure–child*, *co-parents*, and *babysitter–child*, reflecting cultural and social roles that are more explicitly lexicalized in English.

These observations highlight how cultural context shapes the granularity and salience of social relationships represented in dialogue datasets.

Diverse interpersonal dynamics Figure 7 illustrates that our dataset can capture diverse interpersonal dynamics by labeling relational dimensions. The typicality of certain relationships is often defined by their levels of intimacy, formality, and hierarchy (Wish et al., 1981). For instance, friendship is generally characterized as intimate, non-hierarchical (equal), and informal. Yet, our dataset also includes atypical relationships. For instance, over 40% of friend relationships in our dataset deviate from these typical dimensions.

Annotation Type	EN	KO
Hierarchy (All)	0.333	0.462
Hierarchy (2>)	0.416	0.550
Formality (All)	0.408	0.469
Formality (2>)	0.513	0.562
Intimacy (All)	0.314	0.375
Intimacy (2>)	0.426	0.458
HIGHLY-LIKELY (Jaccard)	0.266	0.343
UNLIKELY (Jaccard)	0.837	0.694

Table 6: Inter-annotator agreement for auxiliary relational dimensions and relationship labels. For Hierarchy, Formality, and Intimacy, we report agreement on all annotated samples (*All*) and on samples where at least two annotators selected a non-neutral label (2>). For HIGHLY-LIKELY and UNLIKELY relationship labels, we report mean pairwise Jaccard similarity because annotators may assign multiple labels to a single dialogue.

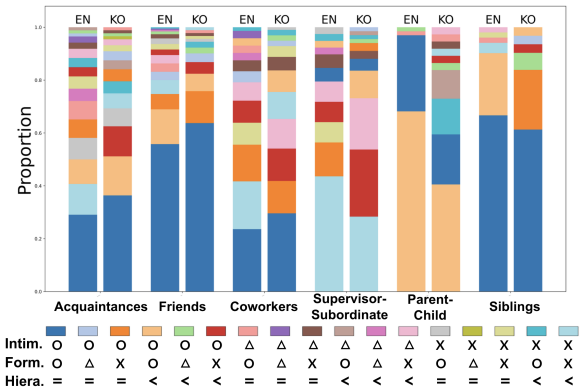


Figure 7: **Comparative analysis of relational dimension distributions for six relationship types present in both English and Korean top-10 relations.** Legend labels denote intimacy (O: intimate, X: unintimate, △: neutral), formality (O: formal, X: informal, △: neutral), and hierarchy (<: hierarchical, =: equal).

B Experiment

B.1 Model Configuration

We evaluate nine models: GPT-4o, o3, Gemini-2.5-Flash, Qwen-3-{8B/14b} (Yang et al., 2025), Llama-3.1-8B-instruct (Grattafiori et al., 2024), A.X-4.0-light-7B, Kanana-8B , and Exaone-4.0-30B (LG AI Research, 2025).[12] For GPT-4o, we use the gpt-4o-2024-05-13 version via the OpenAI API. For Gemini-2.5-Flash, we use OpenRouter (openrouter.ai). We set the temperature of the models to 1.

B.2 Prompt

B.2.1 LLM-as-a-Judge Prompt

LLM-as-a-Judge Prompt

You are a judge that evaluates the correctness of relation classification outputs from the model.
Your task is to check each relation in the model's output and determine whether it belongs to the *Possible* or *Impossible Relation Sets* provided.
For each relation:

- If it is included in the Possible Relation Sets, mark `"is_in_possible": 1`, otherwise mark it as 0.
- Similarly, if it is included in the Impossible Relation Sets, mark `"is_in_impossible": 1`, otherwise mark it as 0.

Answer **only** in the following JSON format (no explanations):

Possible Relation Sets: `{gt_list}`
Impossible Relation Sets:
`{impossible_gt_list}`
Model output: `{model_outputs}`
Output in JSON format:

```
{
    relation1 : {
      "is_in_possible": 1,
      "is_in_impossible": 0
    },
    relation2 : {
      "is_in_possible": 0,
      "is_in_impossible": 1
    },
    ...
}
```

Output:

B.2.2 Inference Prompt

Inference Prompt (EN)

Read the following conversation and guess the relationship of the participants [A] and [B]. When guessing the relationship, refer to the following examples of relationships:
`{example_relations}`
If the relationship matches one of the examples above, use it as is, but if the relationship does not fit any of the examples, describe the relationship yourself.
Your answer about the relationship must be in JSON format:

```
{
  "relation": ""
}
```

`{Additional Information}`
Conversation: `{dialogue}`
Output (JSON):

Inference Prompt (Ko)

다음 대화를 읽고, 등장인물 A와 B의 관계를 추측하시오.
관계를 추측할 때는 다음의 관계 예시를 참고하시오:
`{example_relations}`
만약 위의 예시에 해당하는 관계라면 그대로 사용하고, 예시에 없는 관계라고 판단되면 해당 관계를 직접 서술하시오.
관계에 대한 최종 답은 반드시 JSON 형식으로 답변하시오:

```
{
  "relation": ""
}
```

`{Additional Information}`
대화: `{dialogue}`
Output (JSON):

[12] https://platform.openai.com/docs/models/gpt-4o
https://platform.openai.com/docs/models/o3
https://github.com/SKT-AI/A.X-4.0
https://github.com/kakao/kanana

Example Relations

English Relations

```
[
  "Parent-Children",
  "Brothers/Sisters",
  "Grandparent-Grandchildren",
  "Cousins",
  "Uncle/Aunt-Niece",
  "Friends",
  "Acquaintances",
  "Neighbors",
  "Strangers",
  "Romantic Interest",
  "Dating",
  "Married",
  "Engaged",
  "Friends with benefits",
  "Affair",
  "Ex-relationship",
  "Coworkers",
  "Professional colleagues",
  "Supervisor-Subordinate",
  "Mentor-Mentee",
  "Teacher-Student",
  "Lawyer-Client",
  "Doctor-Patient",
  "Landlord-Tenant",
  "Competitive relationship",
  "Rivalry",
  "Arch-enemies"
]
```

Korean Relations

```
[
  "부모-자식",
  "형제/자매/남매",
  "조부모-손주",
  "사촌",
  "삼촌/이모/고모-조카",
  "단짝 친구",
  "친구",
  "지인",
  "이웃",
  "모르는 사이",
  "썸",
  "연애",
  "부부",
  "약혼관계",
  "Friends with benefits",
  "불륜관계",
  "전애인 관계",
  "동료",
  "직장 동료",
  "상관-부하직원 관계",
  "멘토-멘티",
  "선생-제자",
  "변호사-고객",
  "의사-환자",
  "집주인-세입자",
  "경쟁관계",
  "라이벌 관계",
  "숙적"
]
```

CoT setting: We append *"Think step by step"* at the end of the prompt to encourage chain-of-thought reasoning.

B.2.3 Prompts Used in §6

For the experiment in §6, we add additional information to the prompt. The additional information consists of two types: Age & Gender and Relational Dimensions. The prompts for each type are as follows.

In the Ground Truth Labels setting, we fill the placeholders {age_gender_info} and {relational_dimensions_info} with human-annotated gold labels. In the Model-Generated Labels setting, the model is first asked to separately infer each type of information, and the inferred information is then inserted back into the corresponding placeholders.

Additional Information (En)

Base setting: None

Age & Gender: The age and gender information of the participants [A] and [B] are as follows. Please refer to them when inferring the nature of their relationship. `{age_gender_info}`

Relational Dimensions: The Intimacy level, Pleasure level, and Hierarchy level between A and B in the conversation are as follows. Please refer to them when inferring the nature of their relationship.
`{Relational Dimensions_info}`

Additional Information (Ko)

Base setting: None

Age & Gender: 등장인물 A와 B의 나이와 성별은 다음과 같다. 그들의 관계의 성격을 추론할 때 참고하라. `{age_gender_info}`

Relational Dimensions: 대화에서 A와 B 사이의 친밀감(Intimacy) 수준, 격식(Formality) 수준, 그리고 위계(Hierarchy) 수준은 다음과 같다. 그들의 관계의 성격을 추론할 때 참고하라.
`{Relational Dimensions_info}`

B.3 Validating LLM-as-a-Judge

To validate the accuracy of GPT-4o as an evaluator, we sample 100 question–answer pairs for each language, and two authors independently verify the results. This verification process shows that GPT-4o correctly evaluates 97.85% of the responses in English and 86.2% in Korean, with its judgments matching those of two authors on 96% of the data points (Cohen's $\kappa = 0.58$).

B.4 Results

See Table 7 for the main results, Table 8 for results with CoT prompting, and Table 9 for results from Korean-specialized models.

Model	English		Korean	
	HIGHLY LIKELY (↑)	UNLIKELY (↓)	HIGHLY LIKELY (↑)	UNLIKELY (↓)
GPT-4o	0.791	0.109	0.693	0.219
Gemini-2.5-flash	0.758	0.155	0.589	0.320
Qwen-3-8b	0.565	0.236	0.423	0.335
Llama-3.1-8b	0.413	0.307	0.321	0.522
GPT-4.1	0.819	0.084	0.783	0.241
GPT-5	0.794	0.076	0.790	0.243
Qwen3-32B	0.562	0.228	0.552	0.372
Llama3.3-70B-Instruct	0.525	0.154	0.540	0.381
kanana2-30b-a3b-instruct	0.556	0.258	0.561	0.392

Table 7: Comparison of model performance in English (En) and Korean (Ko) datasets. HIGHLY LIKELY represents the accuracy of the model's majority response being a highly likely response, while UNLIKELY indicates the error rate where the model generate an unlikely response.

Model	En		Ko	
	HIGHLY LIKELY (↑)	UNLIKELY (↓)	HIGHLY LIKELY (↑)	UNLIKELY (↓)
GPT-4o	0.802 (0.011)	0.097 (-0.012)	0.695 (0.002)	0.203 (-0.016)
Gemini-2.5-flash	0.741 (-0.017)	0.127 (-0.029)	0.603 (0.014)	0.201 (-0.119)
Qwen-3-8b	0.688 (0.133)	0.151 (-0.126)	0.497 (0.023)	0.365 (-0.026)
Llama-3.1-8b	0.541 (0.087)	0.267 (-0.041)	0.300 (-0.058)	0.577 (0.031)
Qwen-3-32B	0.694 (0.132)	0.174 (-0.054)	0.626 (0.074)	0.658 (0.286)
Llama-3.3-70B-Instruct	0.620 (0.095)	0.169 (0.015)	0.603 (0.063)	0.536 (0.155)
kanana2-30b-a3b-instruct	0.569 (0.013)	0.235 (-0.023)	0.601 (0.040)	0.614 (0.222)

Table 8: Comparison of model performance with Chain of Thought Prompting across English (En) and Korean (Ko) with deltas in parentheses.

Model	En		Ko	
	HIGHLY LIKELY (↑)	UNLIKELY (↓)	HIGHLY LIKELY (↑)	UNLIKELY (↓)
ax-4.0-light	0.5889	0.1934	0.4674	0.4127
exaone-4.0-32b	0.3178	0.3074	0.4092	0.4674
kanana-1.5-8b	0.4059	0.2884	0.328	0.3739

Table 9: Performance of Korean Specialized models in English and Korean.

C Qualitative Analysis - Cues

C.1 Original Korean Dialogue

Dialogue 2 (Korean):

(...)
[B]: (경례) 어이.
[C]: 왔냐?
[B]: 야. 뭐냐? 이 익사야? 별로 깊어 보이지도 않는데.
[A]: 익사는 아닌 것 같고.
[B]: 그럼 뭐 유기?
[A]: 아하....유기도 아닌 것 같은데. 너가 가서 한번 봐봐. 한번.
[B]: 그럼 뭐야?
[A]: 야. B아. 그 마음 단단히 먹고 봐.
[B]: 장난하나. 에이씨. 자.. 에이씨.

C.2 What cues do LLMs rely on in social relationship reasoning?

To understand how models use and integrate cues to infer social relationships, we conduct a qualitative analysis on their CoT reasoning.

Terms of Address and Reference LLMs frequently leverage terms of address and references as explicit cues to infer social relationships. For instance, when a speaker use terms like *"Daddy"* or *"Professor [B]"*, the models infer family-based or professional relationship. Self-reference also provide valuable information. For example, a speaker introducing themselves as *"Doctor [A]"* signals their professional identity as a medical practitioner, leading to LLMs suggesting relationships such as Doctor-Patient or Doctor-Doctor. Furthermore, LLMs analyze how individuals refer to third parties to understand the relationship between the referring individuals themselves. For example, if both A and B refer to a third person as *"Sergeant [C]"*, the LLM can infer that A and B are likely colleagues within a military context, and that their shared use

of a formal title suggests a potentially task-oriented conversation.

Conversation context and background LLMs also take into account the context of the conversation (e.g., *a school, church, workplace, home*). They then utilize this background information to infer the social relationship or level of intimacy between the individuals involved in the dialogue.

Tone or Atmosphere LLMs also assess the emotional tone of individuals in a dialogue to judge their social dimensions, particularly intimacy and formality, utilizing that information to infer their social relationship. The models often associate *casual, friendly, teasing, empathetic, or supportive* tones with more intimate relationships while *aggressive, frustrated, or angry* expressions are linked to less intimate or strained relationships. Similarly, emotional expressions, whether friendly or hostile, are often connected to informal relationships while the models associate *serious, indifferent, dismissive, or emotionally neutral* expressions with formal relationships.

Relational Dimensions When inferring social relationships, models often consider relational dimensions (*intimacy, hierarchy, formality*) in their rationale. For instance, in a dialogue where A playfully jokes with B while B shares personal concerns, the model infers strong intimacy and suggests a close tie such as friendship or siblinghood.

However, it is important to note that while using social dimensions as cue, particularly hierarchy, LLMs often reveal social stereotypes, attributing *"typical"* relational dimensions to certain relationships. For example, models assume that a parent-child inherently shares a *hierarchical* relationship while a married couple would generally have a *non-hierarchical (equal)* relationship. This leads to reasoning failures when the actual interaction deviates from these norms.

D Does Providing Additional Relational Dimension Help?

This section provides supplementary material for Section 6. Tables 10–11 present the results across models, and Table 12 reports the accuracy of inferring relational dimension.

Associations Between relational dimension and Social Relationship Reasoning Performance To further examine the link between relational dimension inference and relationship reasoning, we run separate logistic regressions for each factor. Table 13 shows that most factors are positively associated with social relationship reasoning. This suggests that models performing well on age, gender, and relational dimension inferences also tend to perform better on overall social relationship reasoning, highlighting the interconnections among these dimensions.

Model	Δ Highly Likely (↑)	Δ UnLikely (↓)
GPT-4o		
Age & Gender	-0.0411	-0.0479
Rel. Dims	-0.0975	-0.0047
Both	-0.0754	-0.0121
Gemini-2.5-flash		
Age & Gender	-0.0411	-0.0343
Rel. Dims	-0.1563	-0.0570
Both	-0.1452	-0.0534
Qwen-3-8b		
Age & Gender	0.0127	0.0392
Rel. Dims	-0.1580	0.1209
Both	-0.1138	0.0398
Llama-3.1-8b		
Age & Gender	0.0205	-0.1027
Rel. Dims	-0.0123	-0.0735
Both	0.0505	-0.0994

(a) With Ground Truth Labels

Model	Δ Highly Likely (↑)	Δ UnLikely (↓)
GPT		
Age & Gender	-0.1507	-0.0137
Sub Dims	-0.1344	-0.0084
Both	-0.1529	0.0285
Gemini		
Age & Gender	-0.1027	-0.0137
Sub Dims	-0.1194	-0.0460
Both	-0.1711	-0.0017
Qwen		
Age & Gender	-0.0395	0.0321
Sub Dims	-0.0237	0.0442
Both	-0.0916	0.0435
Llama		
Age & Gender	0.0548	-0.0137
Sub Dims	0.0099	-0.0514
Both	-0.0492	-0.0145

(b) With Model-Generated Labels

Table 10: Impact of Relational Information on Model Performance (English).

Model	Δ Highly Likely (↑)	Δ UnLikely (↓)
GPT-4o		
Age & Gender	-0.0122	0.0041
Sub Relation	0.0356	-0.0328
Both	0.0383	-0.0191
Gemini-2.5-flash		
Age & Gender	0.0123	-0.0285
Sub Relation	0.0795	-0.0740
Both	0.0932	-0.0904
Qwen-3-8b		
Age & Gender	-0.1021	0.1796
Sub Relation	-0.0740	0.0795
Both	-0.0631	0.0384
Llama-3.1-8b		
Age & Gender	-0.0285	-0.0735
Sub Relation	-0.0274	-0.1644
Both	0.0082	-0.1315

(a) With Ground Truth Labels

Model	Δ Highly Likely (↑)	Δ UnLikely (↓)
GPT-4o		
Age & Gender	-0.0285	0.0490
Sub Relation	0.0329	0.0028
Both	-0.0055	0.0165
Gemini-2.5-flash		
Age & Gender	0.0123	-0.0244
Sub Relation	-0.0466	0.0795
Both	0.0357	-0.0137
Qwen-3-8b		
Age & Gender	-0.1796	0.1877
Sub Relation	-0.0055	0.0329
Both	-0.0905	0.1014
Llama-3.1-8b		
Age & Gender	-0.1061	-0.0245
Sub Relation	-0.0932	0.1315
Both	-0.0165	-0.1342

(b) With Model-Generated Labels

Table 11: Impact of Relational Information on Model Performance (Korean).

Model	Age	Gender	Intimacy	Formality	Hierarchy	Overall
GPT-4o	49.1%	**57.85%**	62.7%	73.8%	**71.5%**	**60.1%**
Gemini-2.5-Flash	**51.65%**	41.45%	**78.3%**	**76.8%**	69.9%	57.5%
Qwen3-8b	30.25%	44.45%	42.5%	40.0%	54.4%	40.9%
Llama-3.1-8b	42.7%	36.2%	61.6%	71.1%	47.4%	47.8%

Table 12: Accuracy of Inferring Relational Information.

Rel. Info.	Gemini2.5	GPT4o	Llama3.1	Qwen3
Age	0.033	0.001	-0.051	-0.079
Gender	-0.023	0.116	0.074	0.085
Intimacy	0.044	0.095	0.134	0.070
Formality	0.109	0.080	0.099	0.029
Hierarchy	**0.217**	0.164	-0.012	0.150

Table 13: **Regression coefficients for relational dimension inference and social relationship reasoning performance**

Language	Movie Title	Movie Title (Ko)	Genre	Year	# of Scenes
EN	Amelia	-	Adventure, Biography, Drama	2009	17
	Autumn in New York	-	Drama, Romance	2000	26
	Big Fish	-	Adventure, Epic, Drama	2003	27
	Bruce Almighty	-	Comedy, Fantasy	2003	32
	Crazy Love	-	Documentary, Romance	2007	21
	Crazy, Stupid, Love.	-	Romance, Comedy, Drama	2011	39
	Date Night	-	Romance, Comedy, Crime	2010	24
	Easy A	-	Comedy, Drama, Romance	2010	45
	He's Just Not That Into You	-	Romance, Comedy, Drama	2009	16
	Larry Crowne	-	Comedy, Drama, Romance	2011	15
	Monte Carlo	-	Adventure, Comedy, Family	2011	6
	Moonrise Kingdom	-	Romance, Adventure, Comedy	2012	6
	New York Minute	-	Comedy, Adventure, Crime	2004	59
	Something's Gotta Give	-	Comedy, Drama, Romance	2003	21
	Speed Racer	-	Action, Adventure, Comedy	2008	7
	The Blind Side	-	Drama, Biography, Sport	2009	16
	The Bounty Hunter	-	Comedy, Action, Romance	2010	21
	The Brothers Bloom	-	Comedy, Action, Adventure	2008	15
	The Curious Case of Benjamin Button	-	Drama, Fantasy, Romance	2008	22
	The Fault in Our Stars	-	Drama, Romance	2014	17
	The Italian Job	-	Action, Crime, Thriller	2003	22
	The Invention of Lying	-	Comedy, Fantasy, Romance	2009	20
	The Next Three Days	-	Thriller, Action, Drama	2010	12
	The Pacifier	-	Action, Comedy, Drama	2005	16
	The Secret Life of Walter Mitty	-	Adventure, Comedy, Romance	2013	12
	The Theory of Everything	-	Drama, Biography, Romance	2014	13
	Water for Elephants	-	Drama, Romance	2011	15
	Wild Hogs	-	Action, Adventure, Comedy	2007	18
KO	200 Pounds Beauty	미녀는 괴로워	Comedy, Drama, Music	2006	28
	A Violent Prosecutor	검사외전	Action, Comedy, Crime	2016	8
	Battle for Incheon: Operation Chromite	인천상륙작전	Action, Drama, History	2016	9
	Cold Eyes	감시자들	Action, Crime, Thriller	2013	13
	Deranged	연가시	Drama, Sci-Fi, Thriller	2012	15
	Exit	엑시트	Comedy	2019	19
	Extreme Job	극한직업	Comedy, Crime	2019	1
	Hide and Seek	숨바꼭질	Horror, Mystery, Thriller	2013	3
	Jeon Woochi	전우치	Action, Adventure, Comedy	2009	3
	Marathon	말아톤	Biography, Drama, Sport	2005	20
	May 18	화려한 휴가	Drama, History	2007	13
	Miss Granny	수상한 그녀	Comedy, Fantasy, Music	2014	15
	My Tutor Friend	동갑내기 과외하기	Action, Comedy, Romance	2003	36
	Northern Limit Line	연평해전	Drama, War	2015	11
	Ode to My Father	국제시장	Drama, War	2014	11
	Pandora	판도라	Disaster, Action, Drama	2016	29
	Punch	완득이	Comedy, Drama, Sport	2011	25
	Secret Reunion	의형제	Action, Drama, Thriller	2010	24
	Secretly, Greatly	은밀하게 위대하게	Drama, Action, Comedy	2013	13
	Silmido	실미도	Action, Drama	2003	24
	Sunny	써니	Comedy, Drama	2011	31
	Take Off	국가대표	Comedy, Drama, Sport	2009	31
	The Attorney	변호인	Crime, Drama, History	2013	7
	The Berlin File	베를린	Spy, Action, Thriller	2013	18
	The Himalayas	히말라야	Adventure, Biography, Drama	2015	5
	The Neighbors	이웃사람	Thriller, Mystery	2012	36
	The Priests	검은 사제들	Horror, Mystery, Thriller	2015	9
	The Roundup	범죄도시2	Action, Crime, Thriller	2022	21
	The Thieves	도둑들	Action, Comedy, Crime	2012	31
	Tidal Wave	해운대	Action, Drama, Sci-Fi	2009	17
	Tunnel	터널	Disaster, Drama	2016	35
	Veteran	베테랑	Action, Comedy, Crime	2015	6

Table 14: List of movies in **SCRIPTS**. The genre (top three) and release year are sourced from IMDb. The dataset contains 60 movies (English 28 / Korean 32) spanning various genres.