The Shape of Power: A Multilingual Framework for Social Power Reasoning in Dialogues

Farah Atif **Sougata Saha** **Monojit Choudhury**
Mohamed bin Zayed University of Artificial Intelligence
{farah.atif,sougata.saha,monojit.choudhury}@mbzuai.ac.ae

Abstract

Social power plays a fundamental role in shaping human interaction, yet computational studies of power remain limited to narrow linguistic and cultural settings. Existing datasets further lack the demographic and relational depth needed for robust cross-cultural analysis. To address this gap, we introduce a theoretically grounded framework for studying social power in naturalistic multilingual dialogue through movie screenplays. The framework integrates a schema informed by social science theory, a native speaker annotation pipeline refined through pilot studies, and a custom interface for scalable cross-lingual analysis. Using this framework, we constructed an initial corpus containing 15,836 annotated instances from 100 scenes in French and Egyptian Arabic movies. Our analysis reveals strong agreement on observable demographic and contextual attributes, while socially interpretive aspects, such as power asymmetry and intention alignment, remain more contested, highlighting the complexity of social power across cultures. We evaluated 6 Large Language Models (LLMs) and Multimodal LLMs on cross-cultural social power reasoning, finding persistent gaps between human and model agreement in relational and theory-of-mind reasoning. Our work introduces the first extensible multilingual framework for studying social power in dialogues and provides an initial evaluation setting for studying cross-cultural social reasoning.

1 Introduction

Power is a foundational construct in social science; it shapes the structure of societies, governs interpersonal relationships, and is deeply embedded in cultural norms and communicative practices (Foucault, 1980; Hofstede, 2011). From political institutions to everyday conversation, power determines who speaks, who is heard, and whose words carry weight. Because it is simultaneously a social, cultural, and psychological phenomenon, understanding power requires analysis that is both theoretically grounded and empirically broad.

Despite its centrality, power has not been studied systematically across cultures within computational linguistics. Existing work tends to treat power as a monolithic concept tied to a single language or social context, and the conversational corpora available today lack the demographic, relational, and cultural depth needed to study how power is expressed, perceived, and negotiated across societies. As a result, researchers are often constrained to narrow perspectives, specific demographic groups, individual languages, or particular interaction types, which can limit the extent to which findings from one cultural context are validated or generalized to others. This leaves several questions fundamentally open: *i) How is power socially manifested and expressed through demographic and interactional features across cultures? ii) Can LLMs reliably identify power dynamics in naturalistic dialog? iii) How does LLM behavior on power-related tasks vary across languages and cultures, and do multimodal models offer additional gains?* Answering any of these questions is currently intractable, as there is still no principled, scalable way to capture power in naturalistic, cross-cultural interactions, nor sufficiently rich annotated data to support analysis at the required depth.

We address this gap by introducing a framework for the cross-cultural study of social power in naturalistic dialog. The framework consists of three components: i) a theoretically-grounded annotation schema that operationalizes power through demographic, relational, and theory-of-mind variables drawn from established social-science traditions; ii) an annotation pipeline in which trained native speakers produce all annotations, and refined through iterative pilot studies; iii) a custom annotation interface designed around HCI principles for cognitive tractability and cross-lingual extensibility. The framework is language-agnostic

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 19589–19604
October 24-29, 2026 ©2026 Association for Computational Linguistics

by construction, with culturally-specific categories supported where universal taxonomies are insufficient. To demonstrate the framework, we introduce *SocLens*[1]), an initial multilingual dataset spanning two typologically and culturally distinct languages, French and Arabic (Egyptian dialect). *SocLens* contains 100 densely annotated scenes drawn from screenplays grounded in everyday social interactions. Each scene is independently annotated by native speakers, enabling feature-level inter-annotator agreement analysis. We further benchmark state-of-the-art text and multimodal LLMs against the human annotations and analyze their social reasoning behavior.

2 Social Power: A Primer

Social power is one of the most contested concepts in social and political theory (Avelino, 2021), yet scholars across multiple traditions have sought to define it. Marxist accounts root power in control over the means of production, whereby dominant classes shape the state, law, and ideology to reproduce class relations (Althusser, 2024). Weber defines it more broadly as the probability of carrying out one's own will within a social relationship, even against resistance (Weber, 1978) a formulation that later inspired elite and pluralist theories. Against such zero-sum conceptions, Magee et al. (Magee and Galinsky, 2008) argue that power need not operate through domination or resistance; it can equally facilitate cooperation and coordination. Foucault departs further, treating power as relational and diffuse, and embedded in everyday practices rather than concentrated in any class (Foucault, 2012). Although definitions of power remain contested, Raven et al. (FRENCH and RAVEN, 1986) provided one of the most systematic frameworks for understanding social influence by identifying several foundational bases of power. Their model originally proposed five key forms of power: (1) coercive power, which relies on force or threats; (2) reward power, exercised through incentives and benefits; (3) legitimate power, derived from socially accepted authority and norms; (4) expert power, grounded in knowledge and expertise; and (5) referent power, which stems from admiration, identification, or loyalty. Raven later extended this framework by introducing informational power, defined as influence achieved through the control and dissemination of information. Scholars further broadened the concept of power. Mann (Mann, 2012) introduced the notion of ideological power, focusing on the capacity to shape values, beliefs, and perceptions,while Bourdieu's symbolic power (Bourdieu, 2018) captured how power works through naming and classification to confer status and legitimacy.

Power in computational approaches. In computational social science, power has been studied through the lens of language, network structure, and interaction patterns. (Danescu-Niculescu-Mizil et al., 2012) studied how people mirror each other's language and found that people with less power tend to mirror the language patterns of people with more power. (Danescu-Niculescu-Mizil et al., 2013) built a computational model to score politeness using a large-scale corpus of Wikipedia and Stack Exchange requests, showing that powerful individuals tend to be less polite. (Diesner and Carley, 2005) used network analysis to identify executives from email data, finding distinct network signatures associated with high-power positions.

Other work has modeled power more directly, primarily in organizational settings. (Bramsen et al., 2011) inferred hierarchical power relations from linguistic patterns in Enron emails, while (Prabhakaran and Rambow, 2013) studied four forms of power (hierarchical, situational, influence, and power over communication) and their manifestations in written dialog.

However, despite these contributions, two significant gaps remain. First, computational studies have predominantly examined power through linguistic or structural proxies, or within constrained organizational hierarchies, rather than modeling it as a multidimensional social phenomenon situated in broader interpersonal contexts. Second, the majority of existing studies rely on English-language data from Western settings, offering limited insight into how power dynamics operate across cultures and languages. Our framework addresses these limitations by modeling social power through theoretically grounded power types alongside demographic, relational, contextual, and perspective-dependent features, and by extending its study to multilingual, culturally situated interactions.

3 Annotation Framework

3.1 Movies as a Source of Social Power

Movie screenplays have emerged as a valuable resource for studying social phenomena in language (Singh et al., 2022). Unlike synthetic or domain-

[1]https://github.com/mbzuai-nlp/Social-Power-in-Dialogs

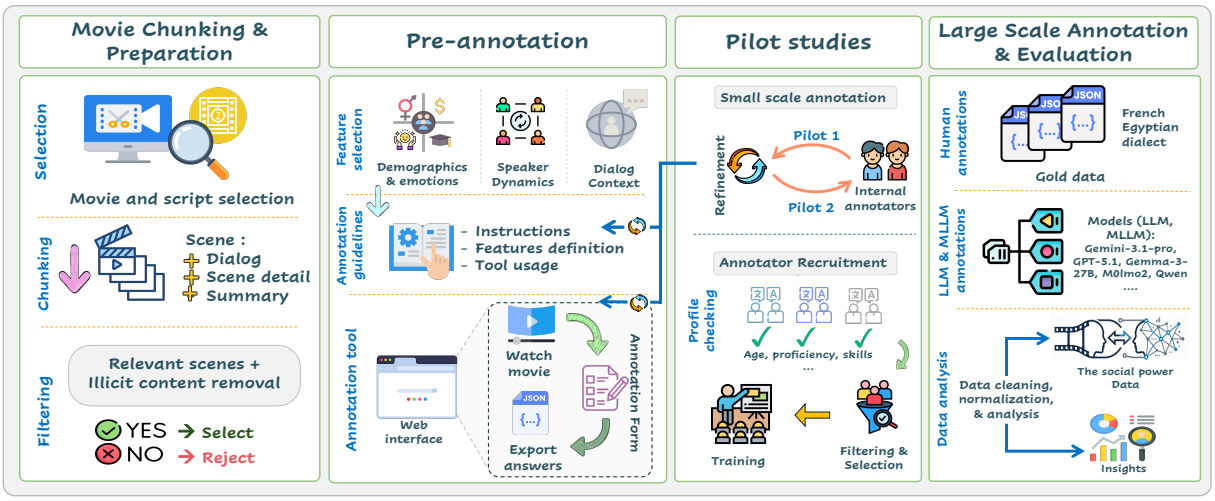

Figure 1: Overview of the annotation pipeline for movie dialogue dataset construction. The workflow comprises four stages: movie chunking and filtering, pre-annotation design (features, tool, and guidelines), pilot studies for annotator training and schema refinement, and large-scale human and LLM/MLLM annotation, followed by cleaning and analysis.

limited corpora (Li et al., 2023), screenplays offer human-crafted, naturalistic dialogue grounded in believable social contexts. They follow standardized formatting conventions that facilitate systematic extraction, and their longitudinal narrative structure allows characters to be traced across interactions (Batty and Baker, 2017), which helps capture how identity, relationships, and power evolve over time. Crucially, films are cultural artifacts; the language and social dynamics of characters reflect the values and interpersonal norms of their cultural context (Adilazuarda et al., 2024), making screenplays particularly suited to cross-cultural analysis of social power.

3.2 Design Principles

The annotation schema and tooling were developed around six guiding principles.

Predictive of power. Every feature in the schema was selected for its theoretical or empirical connection to social power. Social class, occupation, age, gender, relationship type, and goal alignment are all established correlates of power asymmetries in social interaction. Including them together allows the dataset to support both interpretive analysis and predictive modeling of power dynamics.

Annotatable by humans. The schema had to be operationalizable by human annotators with reasonable reliability. This required translating abstract social constructs into observable, dialogue-grounded cues and choosing granularities for categories that balance expressiveness with cognitive tractability. Features that consistently produced low inter-annotator agreement during the pilot were revised or removed.

Assessable by LLMs. In parallel with human annotation, the schema was designed to serve as a benchmark for large language models. Each feature is defined precisely enough to be prompted zero-shot or few-shot, enabling systematic comparison of model and human performance on the same task.

Extendable to different languages and cultures. All features were designed to be language-agnostic and culturally broad. Where universal categories are insufficient, for example, the caste system in South Asian contexts, the schema provides culturally specific alternatives. This ensures the framework can be extended to new languages without structural redesign.

Context and chronology preservation. Power dynamics are relational and cumulative: they cannot be reliably inferred from isolated utterances. The annotation pipeline, therefore, preserves narrative context by running scene summaries and character profiles that evolve throughout the film.

Simplicity for annotators. Both the schema and the interface should be accessible to any annotator reading the guidelines or navigating the tool, without requiring a specialist background. Feature definitions should be concrete and dialogue-grounded, category sets bounded, and the interface should provide a seamless experience that minimizes cognitive load and allows annotators to focus on the

judgment task itself.

3.3 Annotation schema

We annotate five groups of features: demographic attributes of individual speakers, emotional states, power types, dynamic features characterizing the interaction between speakers, and dialogue context. The following subsections describe each feature and its category scheme. Full features and label inventories are provided in Appendix A.1.

Demographic features. We annotate demographic attributes that are theoretically associated with social hierarchy and interpersonal influence, including age group, gender, occupation, educational background, religion, ethnicity, social class, socioeconomic status, marital status, and country of origin. These features provide a socially grounded context for understanding how power is expressed and perceived across cultures.

Emotional State. Social power has also been widely studied in psychology, particularly in relation to the psychological effects of power and social distance (Magee, 2020), as well as traits such as perspective-taking (Galinsky et al., 2016), competitiveness and advice-taking (Tost et al., 2012), and abstract thinking (Magee and Smith, 2013). To capture emotional dimensions, we adopt the six-level emotion taxonomy proposed by (Shaver et al., 1987).

Power Types. We adopt the extended taxonomy of social power introduced by (FRENCH and RAVEN, 1986), and further expand it by including Ideological power.

Speaker Dynamics. Speaker dynamics focus on bilateral relations between interlocutors. For each interaction, we create two directed edges, $SpeakerA \rightarrow SpeakerB$ and $SpeakerB \rightarrow SpeakerA$, since these relationships may be asymmetric. We annotate the following features: perceived power asymmetry, social-status difference, familiarity, relationship type, and goal alignment. Together, these variables capture interaction-level manifestations of power and support the analysis of socially situated reasoning and theory-of-mind inference.

dialogue Context. We annotate contextual properties of the interaction environment, including whether the dialogue occurs in personal or professional settings and whether the interaction takes place in public or private spaces.

4 Dataset Creation

Figure 1 presents the overall workflow used for constructing the dataset. The pipeline consists of four main stages: movie chunking and preparation, pre-annotation design, pilot studies, and large-scale annotation. The process begins with movie and script selection, followed by scene segmentation and filtering of irrelevant or illicit content. Subsequently, annotation features and guidelines are defined and integrated into a web-based annotation tool. Pilot studies are then conducted to refine the annotation schema, train annotators, and improve consistency before scaling to large-scale human and LLM/MLLM-assisted annotation. Finally, the collected annotations undergo data cleaning, normalization, and analysis.

4.1 Movie Selection

Beyond general suitability, we applied a deliberate selection criterion: all chosen films depict recognizable, everyday social situations grounded in human reality. We excluded science-fiction, fantasy, and heavily action-oriented titles in favor of works that emphasize interpersonal relationships, social tensions, and moral dilemmas that audiences across cultures can recognize and relate to. To guide this selection, we consulted native speakers of each target language, asking them to recommend films known for naturalistic dialogue and a strong social or relational focus. We also limited our pool of movies to those available on YouTube with open access to their scenarios.

4.2 Scene Segmentation and Dialogue Extraction

In the first stage, we transform unstructured screenplay text into a structured representation suitable for downstream annotation. Each raw screenplay is segmented into scenes using regular expressions that identify standard scene markers (e.g., `INT.`, `EXT.`, or numbered headings). Each scene is then submitted to an LLM (Gemini-2.5-Pro [2]), which performs three tasks:

Dialogue extraction: the model identifies speaker turns and formats each utterance as `<Speaker: utterance>`.

Scene details: the model produces a concise description of the events, characters, and location of the current scene, which are well-defined in the

[2] https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-pro

screenplay.
Movie summary till now: we prompt the model to summarize all previous scenes until the current one by keeping important events and characters. This summary becomes a memory and helps annotators remember important events.

After formatting the scenes, we perform manual *Scene Filtering and Verification*. We retain scenes relevant to the task and discard transition scenes, scenes with no meaningful events or dialog, and scenes containing explicit content. We also identify scene timestamps and manually verify the extracted dialog against the original screenplay, correcting inconsistencies in speaker attribution, dialog segmentation, and scene mapping. Human verification was conducted iteratively throughout pipeline development to identify recurring extraction errors and refine the preprocessing prompts accordingly. The verified dialogs are then mapped to their corresponding scene and cumulative summaries for downstream annotation.

4.3 Pre-annotation

After identifying the initial set of features to annotate, we developed a custom annotation tool along with an annotation guidelines document. The tool follows Human-Computer Interaction (HCI) best practices, including logical grouping of interface elements (Stureborg et al., 2023), minimizing cross-page navigation (Perry, 2021), and reducing typing burden by offering selectable options rather than free-text fields (Pei et al., 2022; Hollender et al., 2010). The annotation guidelines document provides detailed annotation instructions and interface navigation. Further details about the annotation tool and guidelines are provided in Annotation Guidelines [3].

4.4 Pilot Study

Pilot 1. The first pilot involved three internal annotators and focused on assessing the clarity and feasibility of the annotation task. Annotators worked exclusively with textual inputs, dialogue turns, and a cumulative scene-level summary to label the proposed feature set. The goal was to surface ambiguous or poorly defined features, measure initial inter-annotator agreement, and identify which aspects of the schema required refinement before scaling up.

[3] https://github.com/mbzuai-nlp/Social-Power-in-Dialogs

Pilot 2. Building on the findings of Pilot 1, the second pilot extended the task to a multimodal setting: each scene was linked to its corresponding clip on YouTube, allowing annotators to draw on visual and acoustic cues alongside the screenplay text. In parallel, the annotation guidelines were substantially revised, condensed for readability, supplemented with worked examples, and restructured for easier navigation across sections. This iteration aimed to reduce annotator cognitive load, improve cross-annotator consistency, and evaluate whether access to video led to higher agreement or greater annotation confidence. Across both pilot phases, schema refinement, guideline development, and pilot annotation iterations were conducted over seven months.

4.5 Large Scale Annotation

We conducted a human annotation campaign with native-speaker annotators for each target language. Annotators received the annotation guidelines in advance and attended a practical training session in which each feature was explained in detail, ambiguous cases were discussed, and the interface was demonstrated. Annotators were recruited and paid through Prolific [4]. The large-scale annotation phase was completed over approximately three weeks following the finalization of the annotation schema and pilot studies. We provide further details on recruitment criteria and payment in Appendix A.2 (Annotators recruitment).

Annotation statistics. Table 1 summarizes the corpus statistics. An *annotated item* corresponds to an individual feature annotation. The difference in scene counts results from the filtering process in Section 4.2, with more Arabic scenes satisfying our selection criteria. Despite having fewer scenes and utterances, the French subset contains more annotated items because its selected scenes involve more interacting characters, generating more speaker profiles and directed speaker-dynamics edges.

Language	Scenes	Utterances	Annotated Items
French	38	824	7998
Arabic	62	1027	7838

Table 1: Corpus statistics

5 Evaluation

We conduct two lines of evaluation: i) an inter-annotator agreement (IAA) study to validate the

[4] https://www.prolific.com/

Lang.	Evaluator pair	Age group	Gender	Marital status	Religion	Educ. tier	Occup. tier	SEC	Social class	Emotions	Rel. category	Relationship	Familiarity	Intent align.	Power diff.	Power diff. persp.	Powers	Social diff.	Social diff. persp.	Intent persp.	Domain	Privacy
French	Gemini-3.1-Pro	0.57	0.86	0.46	**1.00**	0.98	0.54	0.58	0.57	0.54	0.82	0.74	0.74	**0.70**	0.32	0.32	0.50	0.40	0.41	0.61	0.47	0.66
	GPT-5.1	0.49	**0.98**	0.39	**1.00**	0.93	0.54	0.40	0.38	0.54	0.46	0.63	0.70	0.62	0.34	0.34	0.29	0.12	0.15	0.32	0.48	0.24
	Qwen-2.5-14B ◇	0.55	0.97	0.38	1.00	0.61	0.55	0.32	0.31	0.35	0.43	0.52	0.53	0.42	0.20	0.18	0.33	0.17	0.17	0.38	0.22	0.40
	Gemma-3-27B ◇	0.59	0.71	0.37	1.00	0.76	0.51	0.49	0.48	0.48	0.53	0.62	0.53	0.38	0.35	0.21	0.25	0.27	0.21	0.06	0.53	0.45
	Molmo2*◇	0.50	0.77	0.23	**1.00**	0.72	0.50	0.35	0.34	0.33	0.41	0.50	0.61	0.34	0.12	0.18	0.32	0.09	-0.02	0.40	0.52	0.56
	Gemini-3.1-Pro*◇	0.74	0.98	0.36	1.00	0.96	0.65	0.56	0.56	0.54	0.84	0.77	0.80	0.69	0.33	0.37	0.62	0.44	0.51	0.52	0.52	0.68
	Human	**0.78**	**0.98**	**0.64**	**1.00**	**1.00**	**0.85**	**0.80**	**0.78**	**0.61**	**0.85**	**0.78**	0.76	0.65	**0.57**	**0.57**	0.59	**0.53**	**0.57**	**0.94**	**0.90**	**0.81**
Arabic-Eg.	Gemini-3.1-Pro	0.53	**1.00**	0.47	0.53	0.60	0.63	0.81	0.81	0.43	0.92	0.83	0.77	0.57	0.23	0.28	0.36	0.71	0.64	0.49	0.71	0.58
	GPT-5.1	0.71	**1.00**	0.53	0.24	**0.67**	0.58	0.57	0.55	0.47	0.85	0.83	0.73	0.48	0.12	0.19	0.44	0.55	0.49	0.17	0.69	0.58
	Qwen-2.5-14B◇	0.47	0.99	0.33	0.19	0.64	0.60	0.51	0.51	0.33	0.81	0.73	0.66	0.44	0.08	0.15	0.37	0.40	0.39	0.08	0.43	0.33
	Gemma-3-27B◇	0.72	0.98	0.39	0.40	0.46	0.53	0.71	0.68	0.40	0.81	0.71	0.74	0.32	0.14	0.14	0.14	0.51	0.40	0.33	0.58	0.45
	Molmo2*◇	0.43	0.97	0.39	0.34	0.39	0.40	0.67	0.65	0.33	0.04	0.39	0.42	0.03	0.17	-0.24	0.51	0.44	-0.22	0.27	0.55	0.37
	Gemini-3.1-Pro*◇	**0.81**	1.00	0.52	0.41	0.61	0.56	0.91	0.89	0.45	0.92	0.83	0.74	0.59	0.20	0.19	0.41	0.64	0.60	0.39	0.73	0.68
	Human	0.62	**1.00**	**0.86**	**0.85**	0.52	**0.77**	**0.93**	**0.90**	**0.68**	**0.97**	**0.85**	**0.83**	**0.69**	**0.44**	**0.38**	**0.59**	**0.83**	**0.78**	**0.93**	**0.89**	**0.68**

Table 2: Inter-annotator agreement scores across evaluator pairs and dimensions for French and Egyptian Arabic datasets.(*) refer to Multimodal LLM's; (◇) indicates models with incomplete coverage, whose agreement scores are computed only over successfully annotated items. Ethnicity and country features were discarded due to perfect agreement across all.

quality and reliability of the human annotations, and ii) a benchmarking study comparing six models against the human annotations across all annotation tiers.

5.1 Inter-Annotator Agreement

Setup. All movies were annotated by two independent native-speaker annotators, yielding parallel annotations for every task. We compute agreement separately for each feature and report Gwet's AC2 (Gwet, 2014) for categorical and ordinal features; a statistic that is more robust than Cohen's κ under skewed distributions (Wongpakaran et al., 2013). For multi-label features such as power types and emotions, we report the mean Jaccard similarity over annotation pairs, as each annotator may assign any subset of labels.

5.2 LLM and MLLM Benchmarking

We benchmark a range of state-of-the-art LLMs and MLLMs. For text-only LLMs, we evaluate two closed-weight models, Gemini-3.1-Pro [5] and GPT-5.1 [6], and two open-weight models, Gemma-3-27b [7] and Qwen-2.5-14B [8]. For multimodal evaluation, we use Gemini-3.1-Pro with video input [9] and Molmo2 (Clark et al., 2026), allowing us to assess whether access to visual context improves power identification beyond what is available from dialogues alone. Each model receives the dialog, the scene summary, the running summary of prior scenes, and the full annotation schema, including feature definitions and permitted value sets (see Appendix A.3 for the prompts). We used an NVIDIA RTX 6000 PRO machine to run the experiments. As for GPT-5.1 and Gemini models, the cost is estimated at approximately 100 USD each. We also fixed the temperature of the models to 0.1. After that, we performed cleaning and normalization steps, including stripping special characters from speaker names, removing incomplete annotations such as missing edges in speaker dynamics, and discarding hallucinated profiles. Hallucinated profiles were defined as generated profiles or speaker-dynamics edges for characters absent from the scene dialog and were automatically detected by matching generated names against the scene speaker list. Profiles with no matching speaker or with the speaker name set to "NA" were discarded, as they could not be reliably mapped to a character.

6 Analysis

6.1 Human and LLM Comparison

Overall patterns. Table 2 reports agreement between human annotators and six models across the French and Arabic-Egyptian subsets. Agreement varies substantially by feature type. Among the

[5] https://deepmind.google/models/gemini/pro/
[6] https://openai.com/index/gpt-5-1/
[7] https://deepmind.google/models/gemma/gemma-3/
[8] https://qwen.ai/blog?id=qwen2.5
[9] https://deepmind.google/models/gemini/pro/

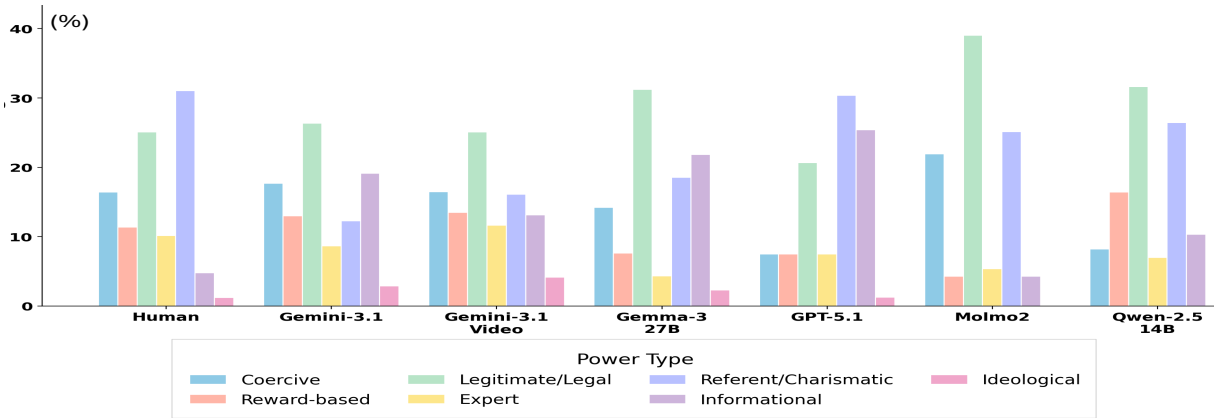

Figure 2: Power type distribution across models and human annotators

text-only models, performance is strongest on relatively observable attributes such as gender, relationship category, familiarity, and educational tier. Gender agreement approaches or reaches 1.00 across most systems, while relationship-related variables also achieve comparatively high agreement. These results suggest that models perform better when social cues are explicit or directly recoverable from the dialog and its context.

Human annotators similarly agree most on demographic and contextual variables. In the Arabic-Egyptian subset, agreement is particularly high for socio-economic class (0.93), social class (0.90), relationship category (0.97), and familiarity (0.83). By contrast, agreement is substantially lower for power difference, social-status asymmetry, and intention alignment, which require relational inference and perspective-taking. Such disagreement may therefore reflect genuine variation in perceptions of social power rather than annotation noise.

Human–model comparisons reinforce this distinction. Models approach human agreement on several demographic and contextual attributes but perform substantially worse on emotions, speaker dynamics, and socially interpretive variables. In Egyptian Arabic, religion agreement ranges from 0.19 to 0.53 among text models, largely because models infer religious identity from culturally common expressions such as *Inshallah* (God willing) or references to Allah (God), whereas annotators often assign "NA" without sufficient evidence. This illustrates how culturally shared linguistic cues can be overgeneralized into demographic attributes. Similarly, GPT-5.1 achieves only 0.12–0.15 agreement on social-status difference in French. Overall, the largest human–model gaps occur on dimensions involving implicit power structures, perspective-taking, and culturally grounded social inference.

Model Coverage. Coverage denotes the proportion of gold items receiving complete, usable annotations. GPT-5.1 and Gemini-3.1-Pro achieve 100% coverage, followed by Gemma-3-27B (92.4%) and Qwen-2.5-14B (89.9%), while both multimodal models cover only 55.7%. Gemini-3.1-Pro (video) frequently skips scenes involving children, likely due to safety refusals, while Molmo2 produces incomplete outputs and unreliable speaker identification. Thus, multimodal agreement scores in Table 2 reflect only successfully annotated items and cannot establish a general benefit from multimodal input.

6.2 Which Power Types Are Most Contested?

Qualitative analysis shows that Referent/Charismatic power is the most contested category, most often in relation to the absence of any identifiable power label. Annotators disagreed on whether Referent power was present in 23% of such cases, compared to only 9% for unidentified-versus-Legitimate distinctions. This suggests that personal influence is less consistently perceived than institutional authority. Among disagreements involving two explicit labels (13% of cases), three recurring patterns emerge. First, Expert versus Referent/Charismatic disagreements (4%) appear only in the Arabic subset, where knowledgeable and charismatic authority often overlap. Second, Legitimate versus Referent disagreements (4%) reflect tension between institutional and personal influence. Third, the most theoretically significant disagreements involve Coercive versus Legitimate power, concentrated in a film centered on an abusive father. One annotator interpreted the character's behavior as paternal authority, while another viewed it as domination and coercion. Rather than reflecting annotation error, these

cases reveal genuine conceptual ambiguity at the boundary between legitimate and coercive authority.

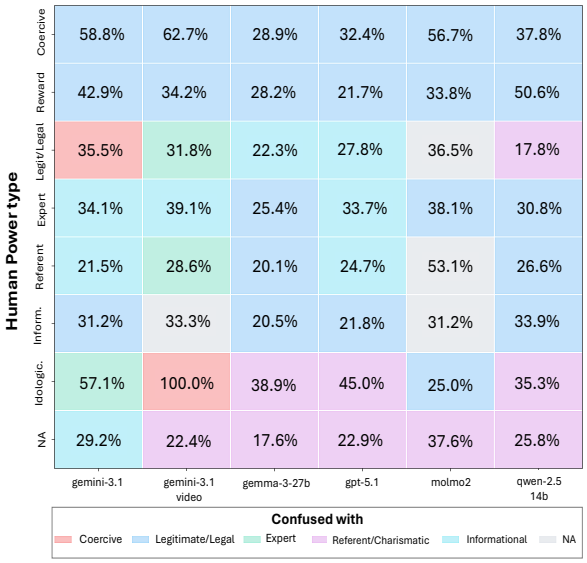

Figure 3: Most frequent human–model power-type confusions across models over the French and Egyptian Arabic data.

Power-Type Confusion Analysis. Figure 3 shows the dominant human–model power-type confusions, where each cell shows the model label most frequently assigned when it differs from the corresponding human power-type annotation. Across models, Coercive and Reward-based power are most frequently collapsed into Legitimate/Legal authority, suggesting that models tend to normalize forms of control into institutionally sanctioned authority. In contrast, Referent/Charismatic and Expert power exhibit greater instability, often being reinterpreted as Informational or Expert authority. Open-weight and multimodal LLMs additionally show stronger collapse into NA labels, particularly Molmo2, indicating difficulties in grounding subtle or socially implicit forms of power even with visual context. Overall, the results suggest that current models recover explicit and institutional forms of authority more reliably than relational or culturally situated forms of influence.

7 Discussion and Conclusion

We introduced a theoretically grounded, multilingual framework for studying social power in naturalistic dialog, comprising a social-science grounded annotation schema, a native-speaker annotation pipeline, and a custom interface designed for cross-lingual extensibility. By positioning social power as the primary object of annotation rather than a downstream proxy, and by capturing relational features directionally from both annotator and character perspectives, the framework enables analysis that prior computational work on power has not supported, including theory-of-mind reasoning, cross-cultural comparison, and fine-grained study of how distinct power bases are perceived and contested. Applying the framework to French and Egyptian Arabic movie dialogues, we find that annotators show strong agreement on observable demographic and contextual variables, whereas more interpretive dimensions, such as power asymmetry and intention alignment, are considerably more contested. Our evaluations further show that although current LLMs handle explicit surface-level social cues reasonably well, they continue to struggle with deeper socially grounded reasoning, especially when cultural understanding and theory-of-mind inference are involved. More broadly, these results contribute to ongoing efforts in computational social science that employ LLMs as proxies for human social behavior and interaction. While such models can reproduce certain patterns of social reasoning, they remain limited in capturing nuanced and culturally embedded power dynamics. This positions social power as a valuable diagnostic for evaluating cross-cultural social reasoning in AI systems, and motivates broader instantiations of the framework across languages, cultures, and modalities.

8 Limitations

This work represents an initial step toward multilingual and cross-cultural modeling of social power, but several limitations remain. First, the current dataset is limited to two languages, French and Egyptian Arabic, which restricts the generalizability of the findings across broader cultural settings. Therefore, we do not claim or assume that certain power types are associated with the cultures or draw any conclusions in this regard. A larger-scale release covering additional languages and substantially more dialogues is planned as a follow-up to this work, and will enable stronger statistical claims about cross-cultural patterns.

Second, while annotators were native speakers of each target language and culturally familiar with the depicted contexts, social power is ultimately a perceptual and interpretive construct. Two annota-

tors per scene, although standard practice, cannot fully capture the range of plausible interpretations of contested dimensions such as power asymmetry or intention alignment. A larger annotator pool per item would allow a deeper study of perceptual variation and its demographic correlates.

Third, our benchmarking covers six state-of-the-art models, but the LLM and multimodal landscape evolve rapidly, and the specific numerical results reported here will date quickly. We mitigate this by releasing the schema, prompts, and annotated data so that future systems can be evaluated under identical conditions. Additionally, our multimodal evaluation is constrained by safety-aligned refusal and structural output failures in some systems, which reduce effective coverage and should be considered when interpreting agreement scores for those models.

9 Ethical Considerations

This work studies socially sensitive attributes and interpersonal power relations, including demographic characteristics, social class, religion, and hierarchy. To mitigate risks of harmful stereotyping or demographic overgeneralization, annotations were performed by trained native speakers familiar with the relevant cultural contexts, and the annotation guidelines explicitly encouraged annotators to avoid unsupported assumptions. Nevertheless, both human annotators and LLMs may still reproduce cultural biases present in media representations and societal norms.

The dataset is constructed from publicly available movie scripts and associated film content intended for public consumption. Explicit or harmful scenes were filtered during pre-processing where appropriate. Annotators were compensated for their work and informed in advance that some scenes could contain sensitive or emotionally difficult material. Finally, we emphasize that the framework is intended for research on social reasoning and cross-cultural analysis, not for profiling, surveillance, or automated judgment of individuals or social groups.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Louis Althusser. 2024. Ideology and ideological state apparatuses: Notes towards an investigation. In *New critical writings in political sociology*, pages 299–340. Routledge.

Flor Avelino. 2021. Theories of power and social change. power contestations and their implications for research on social change and innovation. *Journal of Political Power*, 14(3):425–448.

Craig Batty and Dallas J Baker. 2017. Screenwriting as a mode of research, and the screenplay as a research artefact. In *Screen production research: Creative practice as a mode of enquiry*, pages 67–83. Springer.

Pierre Bourdieu. 2018. Distinction a social critique of the judgement of taste. In *Inequality*, pages 287–318. Routledge.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

Matthew A Diemer, Rashmita S Mistry, Martha E Wadsworth, Irene López, and Faye Reimers. 2013. Best practices in conceptualizing and measuring social class in psychological research. *Analyses of Social Issues and Public Policy*, 13(1):77–113.

Jana Diesner and Kathleen M Carley. 2005. Exploration of communications networks from the enron email corpus. In *Proc. of the Workshop on Link Analysis, Counterterrorism and Security, SIAM Intl. Conf. on Data Mining*, pages 3–14.

Michel Foucault. 1980. *Power/Knowledge: Selected Interviews and Other Writings, 1972-1977*. Vintage.

Michel Foucault. 2012. *Discipline and punish: The birth of the prison*. Vintage.

J. R. P. FRENCH and B. RAVEN. 1986. *The Bases of Social Power*, pages 300–318. University of Pittsburgh Press.

Adam D Galinsky, Derek D Rucker, and Joe C Magee. 2016. Power and perspective-taking: A critical examination. *Journal of Experimental Social Psychology*, 67:91–92.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, et al. 2022. World values survey: Round seven-country-pooled datafile version 5.0.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1).

Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in human behavior*, 26(6):1278–1288.

Michael W Kraus and Wendy Berry Mendes. 2014. Sartorial symbols of social class elicit class-consistent behavioral and physiological responses: a dyadic approach. *Journal of Experimental Psychology: General*, 143(6):2330.

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

Joe C. Magee. 2020. Power and social distance. *Current Opinion in Psychology*, 33:33–37. Power, Status and Hierarchy.

Joe C Magee and Adam D Galinsky. 2008. 8 social hierarchy: The self-reinforcing nature of power and status. *Academy of Management annals*, 2(1):351–398.

Joe C Magee and Pamela K Smith. 2013. The social distance theory of power. *Personality and social psychology review*, 17(2):158–186.

Michael Mann. 2012. *The sources of social power: volume 1, a history of power from the beginning to AD 1760*, volume 1. Cambridge university press.

Saad Z Nagi and Omar Nagi. 2011. Stratification and mobility in contemporary egypt. *Population Review*, 50(1).

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 216–224, Nagoya, Japan. Asian Federation of Natural Language Processing.

Silke L Schneider. 2022. The classification of education in surveys: a generalized framework for ex-post harmonization. *Quality & Quantity*, 56(3):1829–1866.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi Sultan, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5274–5285, Marseille, France. European Language Resources Association.

Christos H Skiadas and Charilaos Skiadas. 2017. A quantitative method for estimating the human development stages based on the health state function theory and the resulting deterioration process. In *Exploring the Health State of a Population by Dynamic Modeling Methods*, pages 21–42. Springer.

Rickard Stureborg, Bhuwan Dhingra, and Jun Yang. 2023. Interface design for crowdsourcing hierarchical multi-label text annotations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2021. Pride: Predicting relationships in conversations. In *The Conference on Empirical Methods in Natural Language Processing*, pages 4636–4650. ACL.

Leigh Plunkett Tost, Francesca Gino, and Richard P Larrick. 2012. Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational behavior and human decision processes*, 117(1):53–65.

William Lloyd Warner, Marchia Meeker, and Kenneth Eells. 1960. *Social class in America: A manual of procedure for the measurement of social status*. Harper.

Max Weber. 1978. *Economy and society: An outline of interpretive sociology*, volume 2. University of California press.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.

Michael Yeomans, Maurice E Schweitzer, and Alison Wood Brooks. 2022. The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology*, 44:293–302.

A Appendix

This appendix provides supplementary material supporting the proposed social power framework and annotation pipeline. We first present the detailed annotation schema and feature definitions, covering demographic attributes, emotional states, power types, speaker dynamics, and dialogue context. We then describe the annotator recruitment protocol, training process, and qualification procedure used during dataset construction. Finally, we provide the prompts employed throughout the framework, including the movie processing and scene extraction prompts, as well as the social power annotation prompts used during large-scale annotation.

A.1 Detailed Annotation Scheme

The annotation framework models social power as a multidimensional phenomenon expressed through demographic cues, interpersonal relations, contextual grounding, and socially situated reasoning. The schema combines observable attributes, such as age or occupation, with interpretive relational variables, including perceived power asymmetry and intention alignment. As illustrated in Figure 4, the framework is organized into five interconnected components spanning demographic features, emotions, power types, speaker dynamics, and dialogue context.

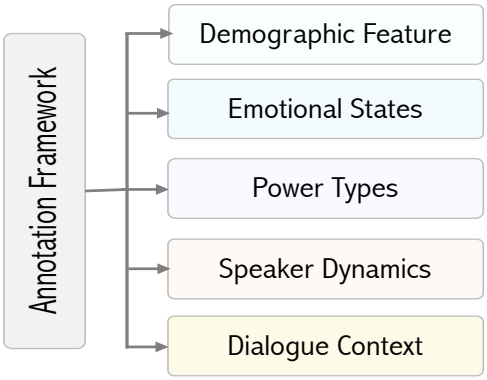

Figure 4: Overview of the proposed annotation framework and its five major components for modeling social power in dialogue.

A.1.1 Demographic features

Age group. Following (Warner et al., 1960) and (Skiadas and Skiadas, 2017), we group age into five broad cohorts: *Child, Teenager, Adult, Middle-aged, Elderly*. Annotators apply their own cultural judgment, as age boundaries vary across cultures.

Gender. Gender is represented through *Male*, *Female*, *Other*, and *NA* categories.

Occupation. We follow the World Values Survey (WVS) coding (Haerpfer et al., 2022), distinguishing employed/self-employed roles (manager, professional, technical, clerical, service, agricultural, craft, operator, elementary, armed forces) from non-employed statuses (retired, student, housewife, unemployed). Occupation is included because professional role and labor position frequently function as visible indicators of institutional authority, expertise, and socio-economic positioning within interaction.

Educational Tier. Following ISCED (Schneider, 2022), we group education into three coarse-grained adaptation of the International Standard Classification of Education (ISCED) (Schneider, 2022). The framework distinguishes *Higher* (Doctoral, Master's, Bachelor's), *Medium* (diploma, secondary), and *Lower* (elementary or below). This coarser grouping reflects the practical difficulty of inferring fine-grained educational attainment from movies, where the fine-grained information is not always easily inferred.

Marital status. Marital status categories are adapted from the World Values Survey coding (), namely: *Married, Single, Divorced, Separated, Widowed, Living together as married* with the addition of *In a relationship / Engaged* to accommodate culturally salient forms of partnership and engagement.

Religion. Religion is represented through broad categories including *Muslim*, *Buddhist*, *Christian*, *Hindu*, *Jewish*, *Atheist*, *Other*, and *NA*. The framework intentionally avoids sectarian distinctions unless they are central to the story where annotators can select "Other" to specify.

Ethnicity. We use a broad regional taxonomy namely: *African / African Descent; Arab / Middle Eastern / North African; Central Asian; South Asian; East Asian; Southeast Asian; Pacific Islander; European / White; Latino / Hispanic; Indigenous / Native Peoples; Mixed / Multi-ethnic; Other; NA.*

Social Class. Following Bourdieu (Bourdieu, 2018), social class is inferred from observable cues such as occupation, speech style, residential setting, clothing, and naming conventions (Kraus and Mendes, 2014; Nagi and Nagi, 2011) and coded as *upper*, *middle*, or *lower*. Annotators are instructed to apply their own cultural framing, as class mark-

ers vary across societies.

Socio-economic Class. We distinguish social class from socio-economic class (SES), as the latter indexes an individual's position within a power hierarchy through relatively objective indicators of resources and capital, such as income, wealth, educational attainment, and occupational prestige (Diemer et al., 2013). The framework distinguishes *Upper*, *Middle*, and *Lower* categories.

Country of origin. The country from which a character originates and is distinct from the location where the dialogue takes place.

A.1.2 Emotional state

The framework adopts the hierarchical taxonomy proposed by Shaver et al. (1987), which organises emotions around six primary categories: *Love*, *Joy*, *Surprise*, *Anger*, *Sadness*, and *Fear*.

Emotional states are annotated at the scene level rather than as stable character traits. The inclusion of affective information reflects the close relationship between emotion, persuasion, conflict escalation, social dominance, and interpersonal alignment.

A.1.3 Power Types

The framework operationalizes power using an extended adaptation of French and Raven's taxonomy of social power (FRENCH and RAVEN, 1986). The schema includes coercive, reward-based, legitimate, expert, referent, informational, and ideological forms of power.

This multidimensional operationalization reflects the view that social power cannot be reduced to institutional authority alone. Instead, power may emerge from expertise, social prestige, ideological influence, charisma, control of information, or the ability to impose sanctions. The framework further assumes that multiple forms of power may coexist within a single interaction for a single character and may vary across cultural settings.

A.1.4 Speakers Dynamics

Power difference. Power difference captures the relative relationship between interlocutors within a given interaction, namely: *Higher*, *Equal*, or *Lower* power, with *NA* reserved for unclear cases. The framework models this relation directionally because social influence is frequently asymmetric. Judgements are collected both from the annotator's global interpretation and from the inferred perspective of individual interlocutors. This distinction allows the framework to capture divergences between annotators and character perception, thereby supporting the study of theory-of-mind reasoning in dialogues.

Social status difference. Social-status difference represents perceived asymmetries in prestige, hierarchy, or social standing between interlocutors. The framework distinguishes *High*, *Equal*, and *Lower* social status. Similarly to power difference, judgements are collected both from the annotator's perspective and the character's perspective.

Familiarity

Familiarity operationalises the degree of social closeness between interlocutors. The feature draws on traditions in sociolinguistics and politeness theory that treat social distance as a major determinant of interactional behaviour, accommodation, and communicative style. Two categories are considered *Familiar* and *Unfamiliar*.

Interlocutor Goal Alignment. Intention alignment captures whether interlocutors pursue *aligned*, *complementary*, or *conflicting* goals within the interaction.

The feature is motivated by recent work on conversational motives and social reasoning (Yeomans et al., 2022). Intention alignment is annotated both from the characters perspectives and from the annotator's broader narrative perspective, enabling analysis of misunderstanding, deception, and socially situated inference.

Relationship Category and Type. Following (Tigunova et al., 2021), relationships are categorized into three groups: *Family* (e.g., parent, spouse, sibling, child), *Social* (e.g., friend, neighbour, enemy, fan), and *Professional* (e.g., colleague, doctor–patient, teacher–student). This feature captures structurally salient interpersonal ties that shape expectations surrounding authority, obligation, familiarity, and emotional expression.

A.1.5 Dialogue Context

The framework additionally annotates contextual properties of the interaction environment. **Location domain** distinguishes between *professional* and *personal* settings, while **location privacy** captures whether the interaction occurs in *public* or *private* space.

These contextual variables are included because social behaviour and power expression are subject to situational environment, audience presence, and institutional setting.

Movie preprocessing

```
You are a movie script processing assistant.  You are given a chunk of a movie
script.Your task has three parts:

1.  Extract Dialogue
- Make sure to group dialogues that belong to the same scene, think step by step when
grouping dialogues.
- Only extract lines of actual dialogue.
- Format:  Each line must follow this exact format:  "CHARACTER: dialogue".
- Ignore all descriptions, scene directions, or narration.
- If the line contains any **illicit or harmful content** (e.g., self-harm, abuse,
jailbreak), **replace that entire line** with:  "<ILLICIT>".
- Do not merge dialogues that do not belong to the same scene.  Split wisely and when
necessary.
- The scenes start usually with examples Therefore, split accordingly.
- If there is no dialogue in a scene no need to include it.
--
2.  Build Memory of Events and Characters

1.  Create a summary of the current events from the provided script
2.  Consider the past events in the {memory}
3.  Merge the two summaries to create a coherent memory that describes all important
events that happened.  A user who is not familiar with the movie should be familiar
after reading the summary.
- Each memory should provide a brief description of the persons involved in the scene
like (Age, occupation, education, ...etc).
- The summary should include Before events and Current events.  In before describe major
events in previous scenes.  And now, what is currently happening in this scene.
- Do NOT reset memory each time.  Think of memory like a snowball:  each group adds to
it.
- The final memory should feel like an **ongoing, concise summary of everything important
so far** (including both past and new information).  - The summary should be in the same
language as the original script in {language}.
- Refine the memory by incorporating new events and removing less relevant ones.
- Make sure that the summary will help the reader understand the movie no matter from
where they start reading it.

3.  Scene specific details
- For each dialogue include a description of the people, the scene, and relevant
information to understand the current dialogue.

**Input Memory**:
{memory}

**Script Chunk**:
{dialogues}

--
Output Format
- **dialogues**:  a list of dialogue lines from that group, formatted as "CHARACTER:
dialogue".
- **scene details**:  (A concise description of the current scene ) short around 200
words.
- **overall summary**:  (The comprehensive, cumulative summary of the entire story so
far) around 400 words.
- The output must be a JSON.
```

A.2 Annotators recruitment

We recruited four annotators, two per language, through the Prolific platform, subject to the following criteria: (i) native speakers of the target language with strong familiarity with the associated cultural norms and conventions; (ii) aged 18 or older; (iii) holding at least a Bachelor's degree, with proficient reading and writing competence in English; and (iv) they must possess a Prolific account. All annotators completed a two-hour online training session, followed by two qualification tasks designed to assess their understanding of the schema before proceeding to the main annotation. Annotators were compensated at a rate of 12 USD per hour, with compensation also covering time spent watching the movies and reviewing the annotation guidelines.

A.3 Prompts

In the following section, we present the "Movie Processing Prompt" used during the movie preprocessing and scene extraction stages, followed by the "Annotation Prompt" which was carefully designed to capture the nuanced aspects of social power during the annotation stage of the proposed framework.

Annotation Prompt

You are a movie script processing assistant. You are given a scene from a movie script, including a dialogue, an overall summary and a scene summary.
Disclaimer: Some of the scenes may contain sensitive content, arguments, explicit language. We expect you to be able to process such content and provide your analysis without any issue. You can ignore explicit content as well and focus on the analysis. This processing is intended for research purposes to analyze the representation of different social groups in movies and how they interact with each other. We are not asking you to judge the content but rather to analyze it based on the instructions provided.
Your task has three parts, think step by step:
1. Demographic features
- For each character or person involved in the scene, provide the following:
- **Name**
- **Age group**: Child, teenage, adult, middle age, elderly
- **Gender**: Male, female, other, NA
- **Ethnicity**: 'African / African Descent', 'Arab / Middle Eastern / North African', 'Central Asian',' South Asian (e.g., Indian, Pakistani, Bangladeshi)', 'East Asian (e.g., Chinese, Japanese, Korean)', 'Southeast Asian (e.g., Filipino, Vietnamese, Thai)',' Pacific Islander / Oceanian', 'European / White', 'Latino / Hispanic', 'Indigenous / Native Peoples', 'Mixed / Multi-ethnic', 'Other','NA'.
- **Marital Status**: 'Married','Single','Divorced', 'Separated', 'Widowed', 'Living together as married', 'In a relationship/engaged','NA'
- Educational tier: low education (Elementary or Secondary), medium education (High School, Diploma (technical or vocational)), high education (Bachelor's, Master's, Doctoral'), NA (When the information is unclear)
- **Religion**: 'Buddhist','Christian', 'Hindu', 'Muslim', 'Jew', 'Other religion','Atheist','NA'
- Socio-economic class : Upper, middle, lower, NA if not clear or hard to determine.
- **Social class**: Upper, middle, lower, NA if not clear or hard to determine.
- **Occupation tier**: Employed/self-employed , No/unpaid, NA if hard to determine.
- **Country**: Name of the country, NA when hard to infer, Fictional if the country is fictional. - **Emotions**: Select one or many emotions from the following category: Love, Joy, Anger, Sadness, Fear, Surprise.
- **Power type**: Select one or many power types for this person: 'Coercive', 'Reward-based', 'Legitimate/Legal', 'Expert','Referent/Charismatic', 'Informational', 'Ideological', or 'NA'.
'Coercive': 'Coercive power relies on the ability to punish or enforce consequences for disobedience. E.g: a manager threatening to fire an employee',
'Reward-based': 'Reward power comes from the ability to provide positive incentives or benefits. E.g: a manger offering bonus',
'Legitimate/Legal': 'Legitimate power is based on a person's formal position or authority. E.g: a police officer enforcing the law, a mother making rules for her children',
'Expert': 'Expert power comes from possessing specialized knowledge, skills, or experience that others respect. E.g: a doctor, a teacher explaining a concept',
'Referent/Charismatic': 'is based on personal qualities that make others admire or trust. E.g: a charismatic, a celebrity, a charismatic person',
Informational': 'Informational power arises from controlling access to important data or insights. E.g: a journalist with exclusive information, a person with insider knowledge',
'Ideological': 'Ideological power is rooted in shared beliefs or values that inspire others to act. E.g: a political leader rallying supporters, a religious leader guiding followers.
'NA': 'hard to determine or the dialogue does not exhibit any type of power for a speaker.',
2. Speakers Dynamics
- For each pair of persons you detected earlier.

Annotation Prompt Continued

provide the following:
1. From person A to B:
- **Relationship Category:** Family, social, professional, Other, NA - **Relationship:**
Family: [parent of, child of, spouse of, sibling of, fiancé of, distant family member of, grandparent of, grandchild of, uncle of, aunt of, nephew/niece of, cousin of, Other],
Social: [neighbor of, friend of, lover of, ex-lover of, enemy of, idol of', 'member of same club as', 'Other',],
Professional: ['boss of', 'employee of', 'employer of','teacher of', 'student of','doctor of', 'patient of','seller to', 'client of','colleague of', 'classmate of','religious relationship with', 'Other']
Familiarity (A → B): [Familiar, Unfamiliar, NA]
- **Intentions alignment (from the perspective of A):** [Aligned, Complementary,Conflicting, NA]
Aligned goals occur when both speakers share the same intentions, perspectives, or desired outcomes. Their interaction shows agreement, mutual understanding, or a common purpose.
Complementary goals occur when the speakers' intentions differ but support each other in a productive way. They may bring unique perspectives or roles--such as one generating ideas and the other evaluating them--while still working toward a shared or compatible outcome.
Conflicting goals arise when the speakers' aims or intentions oppose each other. This often happens in arguments, debates, or negotiations where one person's goal contradicts the other's. NA (Not Applicable) applies when the dialogue does not clearly reveal the speakers' goals or when the interaction is neutral, purely informational, or casual. Example: A brief exchange where one speaker gives directions and the other simply acknowledges them.
- **Intentions alignment (from your perspective):** [Aligned, Complementary,Conflicting, NA]
- **Overall power difference "powerDiff" (from the perspective of A):** ['High power', 'Equal power', 'Less power', 'Neutral']
- **Overall power difference "powerDiffPersp" (from your perspective):** ['High power', 'Equal power', 'Less power', 'Neutral']
- **Overall social status difference "socialDiff" (from the perspective of A):** ['Higher status', 'Equal status', 'Lower status', 'Neutral']
- **Overall social status difference "socialDiffPersp" (from your perspective):**['Higher status', 'Equal status', 'Lower status', 'Neutral']
2. You have to do the same, but from person B to A, and the remaining combinations.
3. Intentions alignment For each pair of speakers, determine their intention alignment from your perspective. Intention alignment can be: Aligned, Complementary , Conflicting, NA. Answer NA if it is hard or not possible to determine. Example : "Person A | Person B": "Conflicting"
4. dialogue Context
1. Location domain: [Personal, professional]
2. Location privacy: [Private, Public]
****Overall summary**:**{overall summary}
*Scene description**: scene details
*Input Dialog**: dialog

Instructions
- Understand carefully what the movie is about, the current dialogue and scene. Based on your understanding, return: - **Profiles**: a json list of profiles for the persons involved in the dialogue only. Make sure to only include the persons who are speaking in the dialogue and not the referenced
- **Speakers Dynamics**: a list of json containing speakers dynamics
- **Intention alignment**: a list of pairs of speakers and their alignment e.g: ["Omar, James":"Complementary", "Lee, Samy":"Conflicting"]
- **Dialogue Context**: a dict containing Social setting, Location domain, Location privacy, Dialogue formality.
- The output must be a JSON.