Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges

Bolei Ma[*1], Yuting Li[*2], Wei Zhou[*3,4], Ziwei Gong[*5], Yang Janet Liu[1], Katja Jasinskaja[2], Annemarie Friedrich[3], Julia Hirschberg[5], Frauke Kreuter[1,6], Barbara Plank[1]

[1]LMU Munich & Munich Center for Machine Learning, [2]University of Cologne,
[3]University of Augsburg, [4]Bosch Center for Artificial Intelligence,
[5]Columbia University, [6]University of Maryland, College Park

[*]Equal contributions.

bolei.ma@lmu.de, yuting.li@uni-koeln.de, wei.zhou3@de.bosch.com, zg2272@columbia.edu

Abstract

Understanding pragmatics—the use of language in context—is crucial for developing NLP systems capable of interpreting nuanced language use. Despite recent advances in language technologies, including large language models, evaluating their ability to handle pragmatic phenomena such as implicatures and references remains challenging. To advance pragmatic abilities in models, it is essential to understand current evaluation trends and identify existing limitations. In this survey, we provide a comprehensive review of resources designed for evaluating pragmatic capabilities in NLP, categorizing datasets by the pragmatic phenomena they address. We analyze task designs, data collection methods, evaluation approaches, and their relevance to real-world applications. By examining these resources in the context of modern language models, we highlight emerging trends, challenges, and gaps in existing benchmarks. Our survey aims to clarify the landscape of pragmatic evaluation and guide the development of more comprehensive and targeted benchmarks, ultimately contributing to more nuanced and context-aware NLP models.

1 Introduction

In linguistics, pragmatics studies how context influences the meaning of language (Huang, 2017; Xiang et al., 2024; Birner, 2012), and how people use language in real-life situations to convey implied meanings, emotions, and intentions. Foundational work in this field, such as Grice's (1975) work on implicature and the cooperative principle, Austin's (1975) idea of speech acts, and Sperber and Wilson's (1986) exploration of contextual inference laid the groundwork of the study of language use. These concepts continually influence linguistic studies and also provide insights for computational methods (Mann, 1980; Saygin and Cicekli, 2002; Hovy and Yang, 2021; Cambria, 2025).

However, traditional Natural Language Processing (NLP) models, while competent in syntactic parsing and semantic analysis, struggle with meaning that extends beyond the literal definition of words. This gap is where pragmatics becomes essential. Early approaches, like rule-based systems, use explicitly coded rules for knowledge representation and match input with a knowledge base for response generation (Bajwa and Choudhary, 2006; Grosan et al., 2011). Later, statistical models applied probability theory to predict language behavior (Johnson, 2009), while deep learning, with transformer-based models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), transformed NLP by enabling the generation of contextually relevant text. Yet their performance in understanding pragmatics remains limited (Hu et al., 2023; Sileo et al., 2022; Park et al., 2024b).

Recently, the emergence of Large Language Models (LLMs) has intensified the need to evaluate their human-like communication abilities, particularly for nuanced use cases requiring sophisticated pragmatic reasoning (Hu et al., 2023; Ruis et al., 2023; Yerukola et al., 2024). As LLMs are increasingly deployed in real-world applications, validating their ability to understand and generate contextually appropriate and pragmatically accurate responses is crucial to ensure effective and trustworthy human-computer interactions. While recent advances in LLMs have demonstrated impressive capabilities in generating coherent and contextually appropriate text, their pragmatic competence remains insufficiently evaluated (Hu et al., 2023; Kwon et al., 2023; Chang et al., 2024).

Important questions arise: What resources have been developed to evaluate the pragmatic capabilities of NLP models, and, more importantly,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8679–8696

how can we leverage pragmatics to guide the advancement of LLMs, formalize a comprehensive framework for evaluating pragmatics, and further advance the study of pragmatics in linguistics? To answer these questions, we conduct a comprehensive survey of the resources available for evaluating pragmatics in NLP: In §2 we introduce the core concepts in pragmatics and in §3 we review how these pragmatic phenomena have been evaluated across various NLP tasks. In §4 we summarize how previous works build a pragmatic dataset. In §5 we present the metrics and evaluation techniques used in current research. In §6 we highlight gaps and offer recommendations for future research.

2 Core Concepts in Pragmatics

Pragmatics studies a set of interrelated phenomena and concepts that explain how context influences language interpretation. In this section, we introduce **what** pragmatics is, beginning with those core concepts, we then summarize the pragmatic phenomena examined in the surveyed works.

2.1 Pragmatic Phenomena in Linguistics

We summarize the following fundamental pragmatic phenomena from linguistics (Levinson, 1983; Yule, 1996; Birner, 2012, *inter alia*):

Context. Context is the foundation of interpretating pragmatics, encompassing both linguistic context (e.g., surrounding clauses or sentences) and extra-linguistic context, including temporal, spatial, and social factors (Huang, 2017).

Deixis. In particular, deictic expressions rely heavily on the situational context of an utterance for their interpretation. These expressions include personal deixis (e.g., "I", "you"), spatial deixis (e.g., "this", "that"), and temporal deixis (e.g., "now", "then"), and are interpreted based on factors such as the speaker's location, the time of the utterance, and the prior discourse (Levinson, 2006; Stapleton, 2017; Levinson, 1983).

Implicature. Grice (1975)'s notion of implicature describes how speakers imply additional meanings without stating them explicitly. For example, if Ann asks, "Do you sell paste?" and Bill replies, "I sell rubber cement", the implication is that Bill does not sell paste (Hirschberg, 1985). Implicature results from enriching the context with additional assumptions to comply with the cooperative principle and the conversational maxims: truthfulness (Maxim of Quality), informativeness (Maxim of Quantity), relevance (Maxim of Relevance), and clarity (Maxim of Manner) (Grice, 1975).

Presupposition. Presuppositions are the implicit assumptions that must hold true for an utterance to be meaningful and understood (Stalnaker et al., 1977). For instance, the sentence "The Queen of England is bald" presupposes that England has a unique Queen, even though this fact is not directly stated (Stalnaker et al., 1977).

Speech Acts and Intent Recognition. Understanding speech acts is critical for understanding speaker intent. A speech act is an utterance that not only conveys information but also performs an action (Austin, 1975). Each utterance can be classified as performing an act that fulfills a communicative purpose, such as such as assertion, suggestion or description. For example, the utterance "Can you open the window?" may literally ask about the addressee's ability or, depending on the context, function as an indirect request.

Discourse and Coherence. In linguistics, discourse refers to language use beyond sentence level (Guardado, 2018). Discourse structure is analyzed by examining how sentences and larger text units are interconnected through coherence relations such as elaboration, explanation and contrast. Some theoretical approaches offer hierarchical models that map these relations to reveal the underlying structure of texts (for an overview of Rhetorical Structure Theory, see Mann and Thompson, 1988), while other frameworks provide a formal semantic account that integrates these coherence relations into dynamic discourse representation (for insights into Segmented Discourse Representation Theory, see Asher and Lascarides, 2003).

Social Pragmatics. Social pragmatics broadens traditional theories by exploring how social factors like culture, power, gender, and interpersonal dynamics influence language use. Brown and Levinson (1987) introduced the concept of "face" (self-esteem), showing how politeness strategies manage relationships and preserve social harmony. Gender differences in communication, as examined by Tannen (1990) and Eckert and McConnell-Ginet (1992), highlight contrasting conversational styles between men and women. Cultural variations further complicate communication, as Scollon et al. (2011) demonstrate that politeness norms differ across cultures. In the digital era, social pragmatics also informs computer-mediated communication, where cues like emoticons convey emotions and tone (Herring et al., 2013).

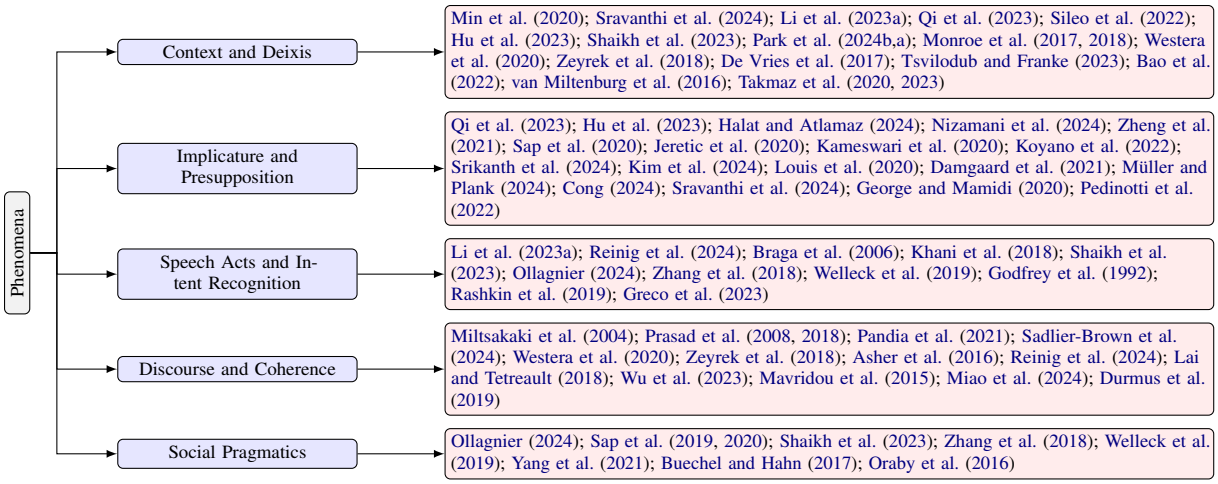

Figure 1: A taxonomy of pragmatic phenomena.

2.2 Pragmatic Phenomena in the Survey

We next examine the pragmatic phenomena addressed in NLP, which were studied in the surveyed works. An overview of the survey methods and inclusivity is presented in Appendix §A. Since pragmatic concepts are interconnected, there are overlaps between categories. Therefore, we focus on the key aspects that each paper addresses. Figure 1 gives an overview of the phenomena and papers.

Context and Deixis. Context and deixis in NLP evaluation tasks usually assess a model's ability to interpret inputs based on the situational or linguistic context. The examination of the context and deixis is the basis of evaluating the models' pragmatic ability as the models rely on the context to respond, similar to the human response process (Greasley and Owen, 2016; van Dijk et al., 2023; Ma et al., 2024a). Datasets or frameworks like Min et al. (2020); Sravanthi et al. (2024); Li et al. (2023a); Qi et al. (2023); Sileo et al. (2022); Hu et al. (2023); Shaikh et al. (2023); Park et al. (2024b,a); De Vries et al. (2017) explicitly address deixis as one of their pragmatic phenomena, requiring models to resolve references in context. In addition, Li et al. (2023a) and Qi et al. (2023) incorporate context-dependent reasoning, requiring models to interpret utterances within multi-turn dialogues. Monroe et al. (2017, 2018) explore language use in color reference tasks, demonstrating how context influences referential choices. Westera et al. (2020) and Zeyrek et al. (2018) also emphasize the role of context in discussion structure and question-under-discussion frameworks. These datasets highlight the role of context in grounding language understanding.

Implicature and Presupposition. A recurring theme across many papers is the study of implicature, i.e., testing the model's inference ability based on the textual input across literal meaning. It explores how meaning is implied rather than explicitly stated, requiring models to infer intentions and assumptions beyond literal interpretations. Different types of implicatures have been studied: Qi et al. (2023), Hu et al. (2023) and Halat and Atlamaz (2024) focus on conversational implicatures, testing models' ability to infer pragmatic meanings in dialogue. Nizamani et al. (2024) investigate scalar implicatures, particularly with gradable adjectives. Zheng et al. (2021) use a grammar-based approach to generate dialogues with intricate implicatures. Cong (2024) looks into LLMs' understanding of manner implicature, a pragmatic inference triggered by a violation of Grice (1975)'s manner maxim. Presupposition is another aspect of implicature and influences the model's inference ability on the implications. Louis et al. (2020); Damgaard et al. (2021); Müller and Plank (2024) focus on indirectness in dialogue. Sap et al. (2020) work on implicatures posted in social media and study social biases. Kameswari et al. (2020) use pragmatic presupposition to enhance bias detection in political news. Moreover, additional works (Jeretic et al., 2020; Koyano et al., 2022; Srikanth et al., 2024; Kim et al., 2024) use pragmatic inferences to understand meanings of the given input.

Speech Acts and Intent Recognition. Speech acts have been explicitly studied in the surveyed works, usually in cases where there are requests, commands, or promises, given to the model and test the model's intent recognition. Li et al. (2023a) include tasks for pragmatic identification and reason-

ing, requiring models to recognize speech acts in multi-turn dialogues. Reinig et al. (2024) provide a fine-grained annotation of speech acts in parliamentary debates. Braga et al. (2006) link prosodic patterns to speech acts and discourse events. A few papers have explicitly different "speakers" involved in the evaluation frameworks and assess how well the models play as the "speakers" in the dialogues, performing the commended task. Khani et al. (2018) explore how players infer intentions and generate pragmatic messages in collaborative tasks, focusing on the rational speech act between the two agents (the "speaker" and the "listener"). Shaikh et al. (2023) involve a multi-turn collaborative two-player game based on the codenames. Ollagnier (2024) also incorporates speech act-like annotations, such as "attack" and "defend", to capture the pragmatic roles of messages in multiparty chats. Both Zhang et al. (2018) and Welleck et al. (2019) talk about the personas, who are the agents making requests, in dialogue and work on the consistency between persona and dialogue.

Discourse and Coherence. Discourse and coherence have been studied in NLP with a focus on parsing text into discourse relations (e.g., Miltsakaki et al., 2004; Prasad et al., 2008, 2018; Miao et al., 2024), on discourse markers (e.g., Pandia et al., 2021; Sadlier-Brown et al., 2024), and on modeling dimensions of coherence (e.g., Lai and Tetreault, 2018; Wu et al., 2023). In addition, discourse structures play a crucial role in dialogue and communication. Westera et al. (2020) and Zeyrek et al. (2018) investigate the role of evoked questions in discourse structure using TED talks. Asher et al. (2016) explore discourse structure in multi-party dialogues and Reinig et al. (2024) study multi-party dialogues in German parliament. Moreover, discourse modes and aspect selection are influenced by contextual and pragmatic factors, connecting coherence to pragmatic interpretation (Mavridou et al., 2015). For a comprehensive overview of situation types and aspects, see Friedrich et al. (2023).

Social Pragmatics. Social pragmatics explores how language use is shaped by social norms, power dynamics, and cultural contexts, including the roles of agents in communication. Ollagnier (2024) introduces a tagset for annotating discursive roles in cyberbullying, while Sap et al. (2019) and Sap et al. (2020) provide benchmarks for reasoning about social norms, biases, and power dynamics. Shaikh et al. (2023) highlight the role of shared cultural knowledge in communication through a collaborative word game, and Zhang et al. (2018) and Welleck et al. (2019) emphasize using personal and social context to improve dialogue systems. Additionally, Yang et al. (2021) examine how pragmatic features vary between neurotypical and neurodiverse speakers, particularly adults with autism. These datasets collectively demonstrate the intersection of social pragmatics with other phenomena like discourse, implicature, and speech acts.

3 Task Types

Understanding pragmatic language use requires evaluating models on a diverse set of tasks that capture various communicative functions and reasoning processes. We then summarize **how** the surveyed works have evaluated pragmatic phenomena, organizing tasks according to general NLP task classifications. They cover a broad spectrum, from structured multiple-choice question answering to open-ended question answering and reasoning, dialogue modeling, and multimodal challenges. A tree graph is shown in Figure 2.

Multiple-Choice Question (MCQ) Setup. MCQs present a scenario or question with predefined answer options, prompting selection of one or more choices. This is commonly used to evaluate LLMs by asking them to choose the correct interpretation of an utterance (Hu et al., 2023; Sravanthi et al., 2024; Park et al., 2024a,b; Kim et al., 2024). It also benchmarks pretrained QA models (Sap et al., 2019; Zheng et al., 2021). Answers can include scalable items or Likert scales (Durmus et al., 2019). MCQs simplify decision-making and improve response accuracy, making them effective for both human studies and LLM evaluations (Groves, 2011; Ma et al., 2024a).

Question Answering (QA). QA, including Conversational QA, models information-seeking dialogues by mapping questions to answers, enabling interactive communication. It evaluates reasoning about unspoken intent, a key aspect of human communication (Grice, 1975). QA tasks typically involve QA pairs (Rashkin et al., 2019; Louis et al., 2020; Srikanth et al., 2024; Kim et al., 2024; Min et al., 2020) or multi-turn dialogues (Li et al., 2023a; Zhang et al., 2018). Some datasets include direct or literal answers to assess models' reasoning abilities (Qi et al., 2023; Miao et al., 2024).

Natural Language Inference (NLI). NLI usually includes a premise and a hypothesis to determine whether the hypothesis is true (entailment),

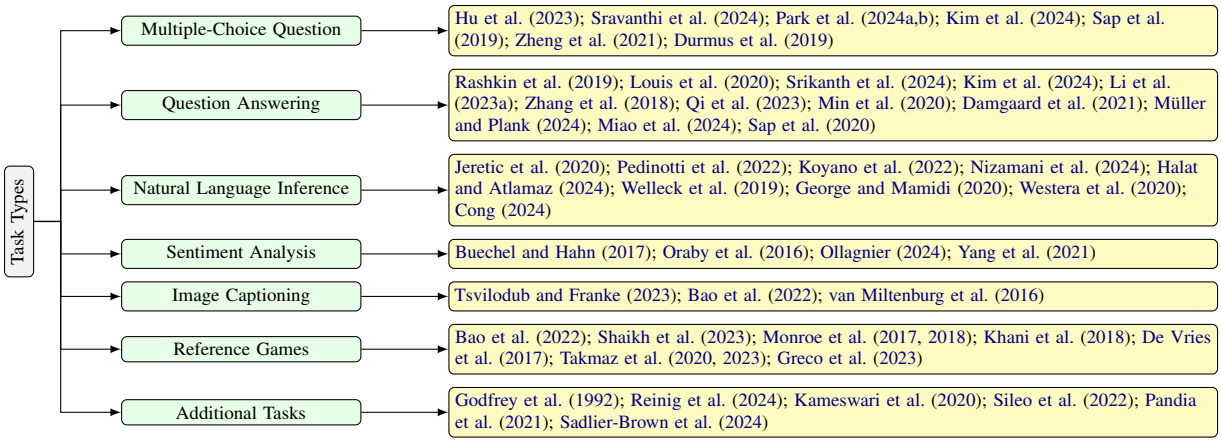

Figure 2: A taxonomy of task types.

false (contradiction), or undetermined (neutral) given the premise (Storks et al., 2020). Pragmatic evaluations often include sentence pairs with certain different pragmatic particles or words (Jeretic et al., 2020; Pedinotti et al., 2022; Koyano et al., 2022; Nizamani et al., 2024; Halat and Atlamaz, 2024) or sentences from dialogues (Welleck et al., 2019), testing whether models can correctly select certain pragmatic (un)related sentences. There are also datasets with NLI-like features looking at the relations of given sentences (George and Mamidi, 2020; Westera et al., 2020; Cong, 2024).

Sentiment Analysis. Sentiment analysis in pragmatics, unlike analysis which focuses on polarity (Gong et al., 2024), extends traditional sentiment classification by incorporating communicative intent, discourse roles, and social interaction dynamics. Datasets in this domain are designed to capture these nuances through fine-grained annotations of conversational exchanges, in topics such as emotion (Buechel and Hahn, 2017), sarcasm (Oraby et al., 2016), hatefulness (Ollagnier, 2024), or medical applications for people with Autism Spectrum Disorder (ASD, Yang et al. 2021).

Image Captioning. Image captioning involves generating descriptions that convey the meaning of a given image. It is a crucial task for studying human reasoning about alternatives and efficient consideration of context (Fried et al., 2023). Unlike other NLP tasks, it is inherently grounded in visual content. Datasets in this area are often designed to investigate pragmatic phenomena, such as contrastive captions and speaker-listener dynamics (Tsvilodub and Franke, 2023; Bao et al., 2022). Some explores negation-based descriptions to enhance pragmatic reasoning (van Miltenburg et al., 2016). These datasets emphasize the interaction between visual context and textual interpretation.

Reference Games. According to Fried et al. (2023), reference games involve a speaker describing a target referent from a shared set of images, objects, or abstract illustrations to a listener, who must then identify the target. This setup is often used to model pragmatic reasoning, such as in tasks requiring context-based contrastive descriptions or multi-turn interactions (Bao et al., 2022; Shaikh et al., 2023). Variants include adaptations of human games to study sequential language use (Monroe et al., 2017; Khani et al., 2018) as well as visual language games in guessing given objects given a visual context (De Vries et al., 2017; Takmaz et al., 2020, 2023; Greco et al., 2023), examining the ability to understand specific referential terms. These datasets emphasize context-dependent interpretation and collaborative communication.

Additional Tasks. There are additional task types designed specifically for certain pragmatic features, including speech-based tasks (Godfrey et al., 1992; Reinig et al., 2024) for speech act studies, bias detection (Kameswari et al., 2020) for assessing social norms, cloze-style tasks for discourse markers (Pandia et al., 2021; Sadlier-Brown et al., 2024), and various natural understanding tasks curated for a pragmatics-centered evaluation framework (Sileo et al., 2022).

4 Dataset Construction

In this section, we summarize methods and data sources used in previous studies for building a pragmatic dataset. Discussing the dataset construction process is crucial for understanding the scope of current pragmatic datasets in terms of domains and

source data, thereby gaining insights about their limitations and special features. We categorize dataset-building approaches for evaluating pragmatic phenomena in more depth, suggesting two groups: a **bottom-up approach**, where source data is first collected and annotation for certain pragmatic phenomena is then performed; and a **top-down approach** which first determines labeling among seed data (e.g., scalar pairs), and then expands seed data to large units, such as sentences, to build up the whole dataset.

4.1 Bottom-up Approach

We discuss this approach in twofold: collecting source data and annotating pragmatic phenomena.

Source Data. In general, there are three types of source data that current pragmatic datasets are built upon: (1) databases, such as web pages and interviews, (2) data collected directly from humans and (3) existing datasets.

There is no one specific **database** for building a pragmatic dataset. Works focusing on dialogue or speech acts use interviews or television programs (Braga et al., 2006; George and Mamidi, 2020) as source data. Common domains for sourcing pragmatic data include open domain (Qi et al., 2023; Wang et al., 2020), debate (Durmus et al., 2019; Oraby et al., 2016), politics (Reinig et al., 2024), social media (Ollagnier, 2024; Sap et al., 2020) and law (Kim et al., 2024). Fewer works **collect data from humans** due to difficulties in recruiting human participants. For instance, Yang et al. (2021) collected spoken language data from adults with high-functioning ASD to examine their pragmatic features. Other works collect data in a game setup, such as reference games, focusing on aspects such as context usage (Khani et al., 2018) and pragmatic inference (Shaikh et al., 2023; Monroe et al., 2017). The majority of datasets are built upon **existing datasets**. They are either a unified benchmark extended on an existing pragmatic dataset (Sravanthi et al., 2024; Sileo et al., 2022; Park et al., 2024b; Guo et al., 2021), or built by annotating non-pragmatic datasets (Li et al., 2023a; Bao et al., 2022; Welleck et al., 2019; Kameswari et al., 2020; Ollagnier, 2024; van Miltenburg, 2017). Conversely, another line of research uses existing datasets not originally related to pragmatics to explore pragmatic phenomena. For example, Ollagnier (2024) adds intent labels to analyze implicature in hate speech.

Annotation. Typical annotation methods involve crowd-sourced annotations (Qi et al., 2023; Li et al., 2023a; Shaikh et al., 2023) and expert annotations (Jeretic et al., 2020; Sravanthi et al., 2024; Hu et al., 2023; Park et al., 2024a). LLMs can also be used to assist annotation. For instance, to create a QA dataset featuring implicature, Srikanth et al. (2024) utilize GPT-3.5 to consolidate a list of assumptions and sub-questions annotated by experts. The model needs to either turn a sub-question into a declarative sentence to create an implicature inference, or identify the assumptions and implicature made in the sub-question. Though speeding up the annotation process and assuring data format, in the post analysis, they find using LLMs to directly generate implicature inference is not reliable. Besides, Cong (2024) employs ChatGPT in generating synthetic stimuli for implicature sentences, with human raters reviewing after the automatic generation process, with the setup of Likert scaling scoring. These hybrid approaches can facilitate the efficient processing of large datasets while maintaining a high level of accuracy through human participation.

4.2 Top-down Approach

Unlike the bottom-up approach, where labels are created after data collection and usually provided by annotators, the labeling process in the top-down approach is motivated by linguistic theories and can be done automatically (Halat and Atlamaz, 2024; Nizamani et al., 2024; Koyano et al., 2022; Sap et al., 2019). For instance, to build an NLI dataset featuring implicature, Halat and Atlamaz (2024) first curate a list of scalar pairs from various linguistic categories. These scalar pairs encode logical relationships naturally, e.g., <*some, all*>. Then they employ GPT-4 to generate sentences based on human-created examples as demonstrations (*I read some of the books.* v.s. *I read all of the books.*). Similarly, to build a pragmatic NLI dataset, Nizamani et al. (2024) collect sentences containing scalar adjectives separated by "*but not*" (e.g., *good but not great*), motivated by the observation that implicatures can be explicitly reinforced (Hirschberg, 1985). To create premise-hypothesis pairs, they split the sentence into two parts, each containing one scalar adjective. For instance, the original sentence: "*These shows were good, but not great.*" can be split into: "*These shows were good.*" and "*These shows were great.*" Manual modification ensures sentence independence and grammatical correctness.

5 Evaluation

We analyze the evaluation of pragmatics in the previous works by summarizing the metrics and methods employed, highlighting their advancements and limitations in capturing pragmatic phenomena.

Evaluation on Model Performance. Currently, automatic metrics dominate the evaluation landscape, with evaluation methods primarily focusing on measuring how well models align with annotated labels or golden data. While this approach is effective for tasks with well-defined categories, it often fails to capture the complexity and subtlety of pragmatic phenomena. For instance, in classification tasks, where average F1 score is widely used, many classification tasks produce binary or categorical labels that reflect only a narrow aspect of pragmatics, such as speech acts (Reinig et al., 2024) or NLI (Nizamani et al., 2024), without addressing broader contextual or interactive dimensions. Similarly, emotion prediction (Buechel and Hahn, 2017) relies on labels (e.g., "joy", "anger") which misses nuanced constructs such as empathy or politeness, and in some applications, the task diverges entirely from pragmatics and uses a binary label for downstream application results. Generation tasks often rely on ROUGE, BLEU, and BERTScore, as seen in cross-cultural pragmatic inference (Shaikh et al., 2023) and social pragmatics (Sap et al., 2020). Other metrics include factuality (Qi et al., 2023), accuracy (Li et al., 2023a; Kim et al., 2024), and task-specific metrics tailored to individual research goals (Li et al., 2023b). While these metrics provide valuable insights into model performance, they often fall short in evaluating the broader, interconnected nature of pragmatics across tasks.

Beyond Automatic Evaluation. Although automatic metrics like F1 score and ROUGE are useful for efficiency and standardization, they can fall short in capturing more nuanced pragmatic aspects, such as politeness. To this end, human evaluation is still needed for more comprehensive evaluation. For instance, in Yang et al. (2021), trained human annotators were asked to rate the scale of politeness and uncertainty in the spoken language data from adults with high-functioning ASD from 1 to 3. These investigated speech features cannot be easily measured by automatic evaluation metrics. Likewise, Rashkin et al. (2019) ran a crowd-sourcing human evaluation on MTurk, where annotators determined whether model-generated responses show empathy. Hence, dialogue models are evaluated not only for its general text generation ability (capturable by automatic metrics), but also more nuanced aspects in human communication, such as emotions. Such a hybrid framework provides a more complete picture of model performance, especially in terms of their capabilities of language uses beyond understanding surface meanings.

We believe it is a promising direction to integrate both automatic and manual evaluation methods that better capture the nuanced and multi-dimensional nature of pragmatic phenomena. Furthermore, there is a growing need for holistic evaluation methodologies that assess a model's ability to handle multiple pragmatic phenomena simultaneously, such as combining emotion detection, politeness, and social reasoning in complex, real-world scenarios (Wu et al., 2024; van Dijk et al., 2023; Sileo et al., 2022). These advancements would enable a more robust understanding of how well models capture the full spectrum of pragmatic behavior. We further discuss this point with an extension to an outlook towards the high-level evaluation frameworks in §6.2.

6 Opportunities and Challenges

We now discuss the opportunities and challenges identified in the surveyed resources, with a focus on the rapid development of LLMs. We focus on generalizability, the high-level evaluation outlook, alignment, as well as the potential of LLMs to advance research in linguistic pragmatics both within and beyond existing resources.

6.1 Gaps and Chances in Generalizability

Based on the surveyed papers, we identify several gaps in the generalizability of current research.

English-Centric Bias. Current pragmatic research is dominated by English-focused datasets, limiting cross-linguistic and cultural generalizability. Out of all 57 surveyed papers, only 11 (i.e., 19%) resources include non-English languages. Though there are a few non-English resources, such as German (Reinig et al., 2024) and Korean (Park et al., 2024a), they focus on regional contexts, lacking critical perspectives for understanding cross-linguistic and cross-cultural pragmatics.

Human/Demographic Bias. NLP frameworks that acknowledge that in certain tasks human labels and judgments may be inherently subjective, and display a lot of variation between subjects (see e.g., Plank, 2022; Cabitza et al., 2023; Kern et al.,

2023). In addition, recently, more and more works have focused on simulating human behaviours and opinions including social demographic profiles of real human participants (e.g., Argyle et al., 2023). The demographics aspect is in general an important factor in the data collection phase in terms of human label variation, and should be considered while conducting LLM evaluations, which we see as a gap in current works.

Data Type Diversity. Most datasets remain largely uni-modal, focusing on either text or speech, with few exploring interactions across modalities and genres. For instance, SWITCHBOARD (Godfrey et al., 1992), while valuable for spoken dialogue, lacks visual data, such as gestures or facial expressions, that are essential for modeling phenomena like deixis, implicature, or politeness. A multimodal dataset combining speech, text, and video could address complex contexts like sarcasm or ambiguity resolution, aligning more closely with real-world communication and advancing pragmatic competence evaluation.

Limitations in Task Types. Pragmatic tasks typically emphasize one or a small set of phenomena, such as speech acts or emotion detection, without capturing the broader interplay between multiple pragmatic dimensions. For example, datasets like SWITCHBOARD (Godfrey et al., 1992) focus on spoken dialogue and discourse markers, while the Lexical Markup Framework (Braga et al., 2006) targets lexical standardization. These tasks are often designed in isolation, making it difficult to evaluate a model's holistic pragmatic competence. Similarly, political discourse datasets like Reinig et al. (2024) analyze specific speech acts but do not integrate additional pragmatic phenomena, such as implicature or politeness. This fragmented approach limits our ability to assess how well models handle the complex, multifaceted nature of pragmatics in real-world scenarios.

Towards Generalizability. Addressing these gaps requires the creation of **multilingual, multimodal datasets** that incorporate visual, textual, and spoken interactions. Leveraging hybrid approaches that combine synthetic LLM-generated data with human validation could provide scalable and diverse benchmarks. In addition, pragmatic task designs should move beyond uni-modal classification and generation to explore holistic, real-world applications that evaluate models' ability to integrate multiple pragmatic phenomena, such as emotion detection, politeness, and implicature resolution, within the same task framework. These advancements will enable a more comprehensive and inclusive understanding of pragmatic competence in computational models.

6.2 Fine-Grained Evaluation of Pragmatics

Despite the growing number of datasets and benchmarks designed to evaluate specific aspects of pragmatic abilities in NLP models, particularly LLMs, there remains a critical need for a comprehensive, high-level framework for fine-grained pragmatic evaluation. Below, we outline key gaps in current evaluation practices and propose recommendations for advancing fine-grained pragmatic evaluation.

Metrics for Pragmatic Alignment. Current evaluation metrics, largely based on traditional NLP tasks, fail to capture the nuances of how well LLMs handle context and social norms. More dynamic and context-sensitive measures, including human-in-the-loop assessments, are needed to assess how well models align with pragmatic expectations. For instance, metrics from psycholinguistics and psychometrics (Shu et al., 2024) are useful to assess the LLM responses.

In-depth Analysis of LLM Outputs. In §3, we drew a diverse landscape of the NLP tasks that the datasets cover. However, most studies treat pragmatics as a classification or generation task, or one of the aspects in their task settings, without closely examining how models arrive at their responses. A more in-depth, qualitative approach—potentially incorporating human feedback and error analysis—could offer deeper insights into where models succeed or fail in pragmatic reasoning. This is particularly relevant for tasks requiring subtle social inferences, implicature resolution, or sensitivity to power dynamics in conversation. We need better measures for evaluating the thought competence in LLMs (van Dijk et al., 2023).

Expanding Data Resources. Unlike syntactic or semantic tasks, pragmatic evaluation relies on rich, contextually grounded datasets, which are difficult to construct at scale. While crowdsourcing and expert annotation are viable solutions, they are resource-intensive. A possible direction is to leverage LLMs themselves to generate controlled datasets, though augmented by human validation (Long et al., 2024). For instance, recent works (Cong, 2024; Srikanth et al., 2024) use GPT to generate synthetic stimuli, with validations from human evaluators.

6.3 Pragmatics Aids LLM Alignment

Recent research emphasizes incorporating linguistic insights into LLM development (Opitz et al., 2025; Brunato, 2025). We also advocate insights from pragmatics for LLM alignment. The alignment of LLMs refers to their ability to generate responses that are not only factually correct but also contextually appropriate and socially coherent (Shen et al., 2023). For instance, speech act theory (Searle, 1969) highlights that meaning emerges from interaction, not isolated sentences—aligning with the communicative goals of LLMs.

Pragmatics in Instruction and Alignment. LLMs rely on prompts to infer user intent, making pragmatics essential for refining instruction-following. Studies have used pragmatic frameworks to evaluate and adjust model responses (e.g., Ruis et al., 2023; Sravanthi et al., 2024; Yerukola et al., 2024), yet they lack robustness in pragmatic inference, failing to recognize indirect requests or subtle shifts in social meaning (Lee and Daniel, 2024). By systematically incorporating pragmatic constraints and examples, LLMs can better interpret ambiguous user inputs and produce responses that align with conversational expectations (Vaduguru et al., 2024).

Human-Computer Interaction. Pragmatics can improve human-AI dialogue and collaboration by ensuring contextually appropriate and socially aware responses. However, chatbot alignment remains limited, especially in handling ambiguity and implicit meaning (Martinenghi et al., 2024). Incorporating pragmatic reasoning can enhance user satisfaction by making interactions more intuitive and responsive to nuanced human intentions, contributing to the continuous advancement of language models (Vaduguru et al., 2024).

Multi-Agent and Collaborative AI. Pragmatics is crucial in multi-agent systems, where agents must infer intent from indirect cues (Li et al., 2024). While a single LLM can role-play different characters with a well-crafted prompt, communication degrades when multiple agents interact, due to LLMs' lack of true belief-state reasoning (Zhou et al., 2024). Addressing this requires models that better handle information asymmetry and implicit intent in multi-agent dialogue.

6.4 LLMs Enhance Research in Pragmatics

Finally, we return to the research in linguistic pragmatics. Pragmatics is a core area of linguistics that has increasingly embraced experimental methods over the past two decades (Sauerland and Schumacher, 2025). Recent advances in LLMs offer new opportunities to advance pragmatic research.

Stimulus and Material Design. Experimental pragmatics relies heavily on carefully designed stimuli to test specific hypotheses (Schwarz, 2017). The recently evolving LLMs might be able to facilitate this by generating diverse and context-rich materials that mirror real-world communication. This can significantly streamline the experimental design process for pragmatic phenomena.

Supporting Data Annotation. High-quality annotation is crucial for pragmatic research but is often costly and time-consuming. LLMs can support the annotation process by providing preliminary labels that human annotators can refine. This hybrid approach has shown promise in previous studies (Tan et al., 2024) but requires careful implementation to ensure annotation quality and reliability (Nasution and Onan, 2024; Rønningstad et al., 2024). Combining LLM-assisted pre-annotation with thorough human review (Wang et al., 2024) offers a more reliable approach to data annotation.

Recommendations for Future Research in Linguistic Pragmatics. Leveraging LLMs' potential in pragmatic research requires collaboration among computational linguists, cognitive scientists, and NLP practitioners. Researchers should explore integrating LLMs into experimental design, material generation, and task development to simulate complex conversational contexts. Rigorous human evaluation and qualitative studies are essential for reliability and interpretability.

7 Conclusion

This survey provides a comprehensive overview of existing resources for evaluating pragmatic understanding in NLP. By categorizing datasets according to the pragmatic phenomena they target and analyzing their task designs with data collection methods and how they are evaluated with the models, we highlight current trends and challenges in this field. Our findings underscore the need for more nuanced and contextually rich evaluation benchmarks, especially as LLMs continue to evolve. We hope to provide a valuable guide for researchers and practitioners advancing pragmatic reasoning in NLP, promoting systems with more human-like communication and interdisciplinary collaboration.

Limitations

We identify the following limitations in the survey.

Survey Sources. Our survey predominantly encompasses literature within NLP, as the scope is to review the resources for evaluating pragmatics in NLP models. Therefore, we did not look extensively for resources in linguistic studies. This might result in an omission of potential resources or enlightening evaluation methods, which could bring insights to current data collection processes and evaluations. We advocate for future surveys in reviewing the resources for pragmatic language understanding in linguistics.

NLP Tasks. Our survey is tailored towards tasks that feature applicability in NLP. As a result, intrinsic pragmatic tasks, such as discourse modeling are not discussed. Nevertheless, they could still be valuable to look into, especially in terms of gaining insights to current evaluation.

Multilinguality. The scope of our survey is confined to literature published only in English. Although we endeavored to include works addressing multilinguality (albeit written in English), it is conceivable that pragmatic datasets created in other languages are not incorporated. This language constraint might limit the inclusivity of diverse pragmatic phenomenons and methodologies that are unique to a specific language.

Multimodality. While our paper covers a few data points that go beyond the textual-based features, such as the image captioning tasks (e.g., Tsvilodub and Franke, 2023), the majority of the data is textual-based. This gap has also been discussed in our §6.1. As we mainly focus on NLP venues, some resources could be not covered. Therefore, we encourage future work to explore pragmatic evaluation across diverse modalities.

Ethical Considerations

In this paper, we examine the evaluation of pragmatics in NLP, particularly in the context of LLMs. Since pragmatic abilities are inherently human, discussing these abilities in LLMs could raise concerns about anthropomorphism, as noted in prior work on LLM evaluation (Ma et al., 2024a,b). It is important to clarify that our discussion of pragmatics does not imply human-like cognition or consciousness in LLMs. Instead, we advocate to use human pragmatic features to build benchmarks, aiming to enhance their design, and to enable user-friendly and effective human-machine interactions.

Moreover, our evaluation focuses on the pragmatic aspects reflected "in" LLMs' textual outputs, rather than attributing beliefs or intentions to the models themselves. This consideration is consistent with recent discussions on LLM evaluation, which emphasize analyzing model outputs without ascribing human-like agency (e.g., Santurkar et al., 2023; Röttger et al., 2024; Durmus et al., 2024). By maintaining this perspective, we aim to contribute to a nuanced understanding of LLMs' capabilities while avoiding the pitfalls of over-personalization.

Acknowledgements

The authors acknowledge the use of ChatGPT exclusively to paraphrase and refine the text in the final manuscript. We thank the members of the MaiNLP lab from LMU Munich for their constructive feedback. YL and KJ are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 281511265 – SFB 1252 "Prominence in Language" in the project C06 at the University of Cologne. ZG is supported by the National Science Foundation via ARNI (The NSF AI Institute for Artificial and Natural Intelligence), under the Columbia 2025 Research Project ("Towards Safe, Robust, Interpretable Dialogue Agents for Democratized Medical Care"). YJL and BP are supported by ERC Consolidator Grant DIALECT (101043235).

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

John Langshaw Austin. 1975. *How to do things with words*. Harvard University Press.

Imran Sarwar Bajwa and Muhammad Abbas Choudhary. 2006. A rule based system for speech language context understanding. *Journal of Donghua University (English Edition) Vol*, 23(06).

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Betty J Birner. 2012. *Introduction to pragmatics*. John Wiley & Sons.

Daniela Braga, Luís Coelho, João P. Teixeira, and Diamantino Freitas. 2006. Progmatica: A prosodic database for European Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. 4. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dominique Brunato. 2025. Learning from impairment: Leveraging insights from clinical linguistics in language modelling research. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4167–4174, Abu Dhabi, UAE. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Erik Cambria. 2025. *Pragmatics Processing*, pages 229–338. Springer Nature Switzerland, Cham.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The one with understanding indirect answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475, Los Alamitos, CA, USA. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Penelope Eckert and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology*, pages 461–490.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational*

Linguistics, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323. Third International Conference on Computing and Network Communications (CoCoNet'19).

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, and Julia Hirschberg. 2024. A mapping on current classifying categories of emotions used in multimodal models for emotion recognition. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28, St. Julians, Malta. Association for Computational Linguistics.

Andrew Greasley and Chris Owen. 2016. Behavior in models: A framework for representing human behavior. In *Behavioral Operational Research: Theory, Methodology and Practice*, pages 47–63, London. Palgrave Macmillan UK.

Claudio Greco, Diksha Bagade, Dieu-Thu Le, and Raffaella Bernardi. 2023. She adapts to her student: An expert pragmatic speaker tailoring her referring expressions to the layman listener. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts: Syntax and Semantics Volume 3*, pages 41–58. Academic Press, New York.

Crina Grosan, Ajith Abraham, Crina Grosan, and Ajith Abraham. 2011. Rule-based expert systems. *Intelligent systems: A modern approach*, pages 149–185.

Robert M. Groves. 2011. Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871.

Martin Guardado. 2018. *What is discourse?*, pages 70–78. De Gruyter Mouton, Berlin, Boston.

Jiaqi Guo, Ziliang Si, Yu Wang, Qian Liu, Ming Fan, Jian-Guang Lou, Zijiang Yang, and Ting Liu. 2021. Chase: A large-scale and pragmatic Chinese dataset for cross-database context-dependent text-to-SQL. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2316–2331, Online. Association for Computational Linguistics.

Mustafa Halat and Ümit Atlamaz. 2024. ImplicaTR: A granular dataset for natural language inference and pragmatic reasoning in Turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 29–41, Bangkok, Thailand and Online. Association for Computational Linguistics.

Susan C Herring, Dieter Stein, and Tuija Virtanen. 2013. Introduction to the pragmatics of computer-mediated communication. *Pragmatics of computer-mediated communication*, pages 3–32.

Julia Hirschberg. 1985. *A theory of scalar implicature*. University of Pennsylvania.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Yan Huang. 2017. Introduction: What is pragmatics? In *The Oxford Handbook of Pragmatics*. Oxford University Press.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Mark Johnson. 2009. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11, Athens, Greece. Association for Computational Linguistics.

Lalitha Kameswari, Dama Sravani, and Radhika Mamidi. 2020. Enhancing bias detection in political news using pragmatic presupposition. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 1–6, Online. Association for Computational Linguistics.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Fereshte Khani, Noah D. Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Kana Koyano, Hitomi Yanaka, Koji Mineshima, and Daisuke Bekki. 2022. Annotating Japanese numeral expressions for a logical and pragmatic inference dataset. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 127–132, Marseille, France. European Language Resources Association.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Bradford J. Lee and Cook R. Daniel. 2024. Exploring the potential of ai for pragmatics instruction. *Technology in Language Teaching & Learning*, 6(3):1521.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge University Press.

Stephen C Levinson. 2006. Deixis. *The handbook of pragmatics*, pages 97–121.

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

Jiaxuan Li, Minxi Yang, Dahua Gao, Wenlong Xu, and Guangming Shi. 2024. Pace: A pragmatic agent for enhancing communication efficiency using large language models. *Preprint*, arXiv:2402.01750.

Yiyuan Li, Rakesh Menon, Sayan Ghosh, and Shashank Srivastava. 2023b. Pragmatic reasoning unlocks quantifier semantics for foundation models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 573–591, Singapore. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024a. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Assenmacher. 2024b. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. *Preprint*, arXiv:2412.13169.

William C Mann. 1980. *Toward a speech act theory for natural language processing*. Information Sciences Institute, University of Southern California.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, and Dimitri Ognibene. 2024. LLMs of catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multiparty dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 107–118, Torino, Italia. ELRA and ICCL.

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sørensen, Alexis Palmer, and Manfred Pinkal. 2015. Linking discourse modes and situation entity types in a cross-linguistic corpus study. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. 2018. Generating bilingual pragmatic color references. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2155–2165, New Orleans, Louisiana. Association for Computational Linguistics.

Christin Müller and Barbara Plank. 2024. IndirectQA: Understanding indirect answers to implicit polar questions in French and Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9025–9035, Torino, Italia. ELRA and ICCL.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795, Torino, Italia. ELRA and ICCL.

Anais Ollagnier. 2024. CyberAgressionAdo-v2: Leveraging pragmatic-level information to decipher online hate in French multiparty chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4287–4298, Torino, Italia. ELRA and ICCL.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing relies on linguistics. *Computational Linguistics*, pages 1–24.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024a. Pragmatic competence evaluation of large language models for the korean language. *Preprint*, arXiv:2403.12675.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024b. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Paolo Pedinotti, Emmanuele Chersoni, Enrico Santus, and Alessandro Lenci. 2022. Pragmatic and logical inferences in NLI systems: The case of conjunction buttressing. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 8–16, Seattle, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Peng Qi, Nina Du, Christopher Manning, and Jing Huang. 2023. PragmatiCQA: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6175–6191, Toronto, Canada. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2024. A GPT among annotators: LLM-based entity-level sentiment annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 133–139, St. Julians, Malta. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 20827–20905. Curran Associates, Inc.

Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. How useful is context, actually? comparing LLMs and humans on discourse marker prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241, Bangkok, Thailand. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Uli Sauerland and Petra B Schumacher. 2025. Pragmatics: two decades of theory and experiment growing together. *Aktuelle Entwicklungen in der linguistischen Forschung*, 35:247.

Ayse Pinar Saygin and Ilyas Cicekli. 2002. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258.

Florian Schwarz. 2017. Experimental pragmatics. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Ron Scollon, Suzanne Wong Scollon, and Rodney H Jones. 2011. *Intercultural communication: A discourse approach*. John Wiley & Sons.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *Preprint*, arXiv:2309.15025.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmatics-centered evaluation framework for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12075–12097, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Robert Stalnaker, Milton K Munitz, and Peter Unger. 1977. Pragmatic presuppositions. In *Proceedings of the Texas conference on performatives, presuppositions, and implicatures. Arlington, VA: Center for Applied Linguistics*, pages 135–148. ERIC.

Andreea Stapleton. 2017. Deixis in modern linguistics. *Essex Student Journal*, 9.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *Preprint*, arXiv:1904.01172.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernandez. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine Books.

Polina Tsvilodub and Michael Franke. 2023. Evaluating pragmatic abilities of image captioners on A3DS. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1277–1285, Toronto, Canada. Association for Computational Linguistics.

Saujas Vaduguru, Daniel Fried, and Yewen Pu. 2024. Generating pragmatic examples to train neural program synthesizers. In *The Twelfth International Conference on Learning Representations*.

Bram van Dijk, Tom Kouwenhoven, Marco Spruit, and Max Johannes van Duijn. 2023. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654, Singapore. Association for Computational Linguistics.

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany. Association for Computational Linguistics.

Lijie Wang, Ao Zhang, Kun Wu, Ke Sun, Zhenghua Li, Hua Wu, Min Zhang, and Haifeng Wang. 2020. DuSQL: A large-scale and pragmatic Chinese text-to-SQL dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6923–6935, Online. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in Chinese

essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688, Singapore. Association for Computational Linguistics.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Mingyou Xiang, Mian Jia, and Xiaohui Bu. 2024. *Introduction to Pragmatics*, volume 9. Springer Nature.

Christine Yang, Duanchen Liu, Qingyun Yang, Zoey Liu, and Emily Prud'hommeaux. 2021. Predicting pragmatic discourse features in the language of adults with autism spectrum disorder. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 284–291, Online. Association for Computational Linguistics.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

George Yule. 1996. *Pragmatics*. Oxford university press.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

A Overview of Surveyed Works

To compile this survey, we conducted a comprehensive review of recent literature on evaluating pragmatics in NLP, with a specific focus on the datasets. We focused on identifying works that address these aspects, using the keywords "pragmatics" and "datasets" by primarily looking at papers publicized by 31.12.2024 at ACL Anthology[1].

This was chosen as the primary source since it is the main publication platform for the *CL community and contains the most relevant works on the evaluation of pragmatics in NLP models. After this initial search, we applied additional filtering criteria to refine the selection. Specifically, we focused on papers that

- introduced datasets explicitly designed for evaluating pragmatic phenomena in NLP,
- discussed methodologies for assessing pragmatic abilities of language models, or
- provided empirical evaluations of NLP models in related tasks to pragmatics.

In addition, we conducted manual qualitative inspections to ensure the inclusion of relevant works that might not have been captured due to variations in terminology; we also included a few recent papers from other sources out of the ACL Anthology based on our expertise in the field. This resulted in a final section of 58 papers in the current version.

B Pragmatic Phenomena and their Corresponding Tasks

In this section, we present the pragmatic phenomena and their corresponding tasks based on the survey works and show their mappings in Table 1. We noticed that there are no 100 percent one-to-one mappings between the pragmatic phenomena and the tasks, i.e., a phenomenon could be evaluated in either task in reality. This shows the possibility of diverse formats to evaluate the pragmatics of LLMs, and calls for future development of finer-grained tasks.

[1] https://aclanthology.org/

Task Type	Context and Deixis	Implicature and Presupposition	Speech Acts and Intent Recognition	Discourse and Coherence	Social Pragmatics
Multiple-Choice Question	Hu et al. (2023); Sravanthi et al. (2024); Park et al. (2024a,b)	Hu et al. (2023); Sravanthi et al. (2024); Zheng et al. (2021); Kim et al. (2024)	–	Durmus et al. (2019)	Sap et al. (2019)
Question Answering	Li et al. (2023a); Qi et al. (2023); Min et al. (2020); Srikanth et al. (2024); Kim et al. (2024); Miao et al. (2024); Sap et al. (2020)	Qi et al. (2023); Louis et al. (2020); Srikanth et al. (2024); Damgaard et al. (2021); Müller and Plank (2024); Kim et al. (2024)	Li et al. (2023a); Zhang et al. (2018); Rashkin et al. (2019)	Miao et al. (2024)	Yang et al. (2021); Sap et al. (2020); Zhang et al. (2018)
Natural Language Inference	Westera et al. (2020)	Jeretic et al. (2020); Halat and Atlamaz (2024); Nizamani et al. (2024); Koyano et al. (2022); Cong (2024); George and Mamidi (2020); Pedinotti et al. (2022)	Welleck et al. (2019)	Westera et al. (2020)	Welleck et al. (2019)
Sentiment Analysis	–	–	Ollagnier (2024)	Yang et al. (2021)	Ollagnier (2024); Yang et al. (2021); Buechel and Hahn (2017); Oraby et al. (2016)
Image Captioning	Tsvilodub and Franke (2023); Bao et al. (2022); van Miltenburg et al. (2016)	–	–	–	–
Reference Games	De Vries et al. (2017); Monroe et al. (2017, 2018); Takmaz et al. (2020, 2023); Shaikh et al. (2023)	–	Shaikh et al. (2023); Khani et al. (2018); Greco et al. (2023)	–	Shaikh et al. (2023)
Additional Tasks	Sileo et al. (2022)	Kameswari et al. (2020)	Godfrey et al. (1992); Reinig et al. (2024)	Pandia et al. (2021); Sadlier-Brown et al. (2024); Reinig et al. (2024)	–

Table 1: Mapping from Pragmatic Phenomena (columns) to NLP Task Types (rows).