Conflicts in Texts: Data, Implications and Challenges

Anonymous ACL submission

Abstract

As NLP models become increasingly integrated into real-world applications, it becomes clear that there is a need to address the fact that models often rely on and generate conflicting information. Conflicts could reflect the complexity of situations, changes that need to be explained and dealt with, difficulties in data annotation, and mistakes in generated outputs. In all cases, disregarding the conflicts in data could result in undesired behaviors of models and undermine NLP models' reliability and trustworthiness. This survey categorizes these conflicts into three key areas: (1) *natural texts on the web*, where factual inconsistencies, subjective biases, and multiple perspectives introduce contradictions; (2) *human-annotated data*, where annotator disagreements, mistakes, and societal biases impact model training; and (3) *model interactions*, where hallucinations and knowledge conflicts emerge during deployment. While prior work has addressed some of these conflicts in isolation, we unify them under the broader concept of *conflicting information*, analyze their implications, and discuss mitigation strategies. We highlight key challenges and future directions for developing conflict-aware NLP systems that can reason over and reconcile conflicting information more effectively.

1 Introduction

The rapid advancement of natural language processing (NLP), particularly with the rise of large language models (LLMs), has led to their widespread adoption in daily tasks, information retrieval, and decision-making processes. However, the increasing complexity of these models reveals various types of conflicts at multiple stages, including training, annotation, and model interaction, affecting the reliability and trustworthiness of downstream applications. For example, training models on data containing factual contradictions, annotation disagreements, or prompts that contradict a model's parametric knowledge can introduce inconsistencies with unpredictable consequences (Pavlick and Kwiatkowski, 2019; Sap et al., 2019).

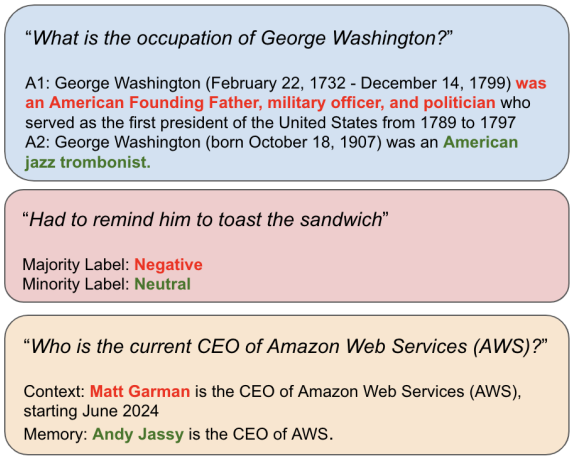

Figure 1: Examples of the three different ares of conflicts discussed in this work. The first example describes a case where two different entities of the same name are found *naturally on the web*, the second example elaborates the *annotation disagreement* in a sentiment analysis task, and the third showcases a knowledge conflict between the context and memory of LLMs during *model interactions*.

Existing work on conflicts in NLP tends to focus on specific issues, such as annotation disagreements (Uma et al., 2021), hallucinations and factuality (Zhang et al., 2023; Wang et al., 2023), and knowledge conflicts (Xu et al., 2024; Feng et al., 2024), without synthesizing these problems into a broader perspective. In this survey, we conceptualize these diverse challenges under the umbrella of *conflicting information* and analyze their origins, implications, and mitigation strategies. We first examine conflicts inherent in training data, including **natural texts from the web** and **human-annotated datasets**. We then explore conflicts arising during **interactions with models** in the LLM era, discussing their impact on downstream tasks and highlighting key challenges and future

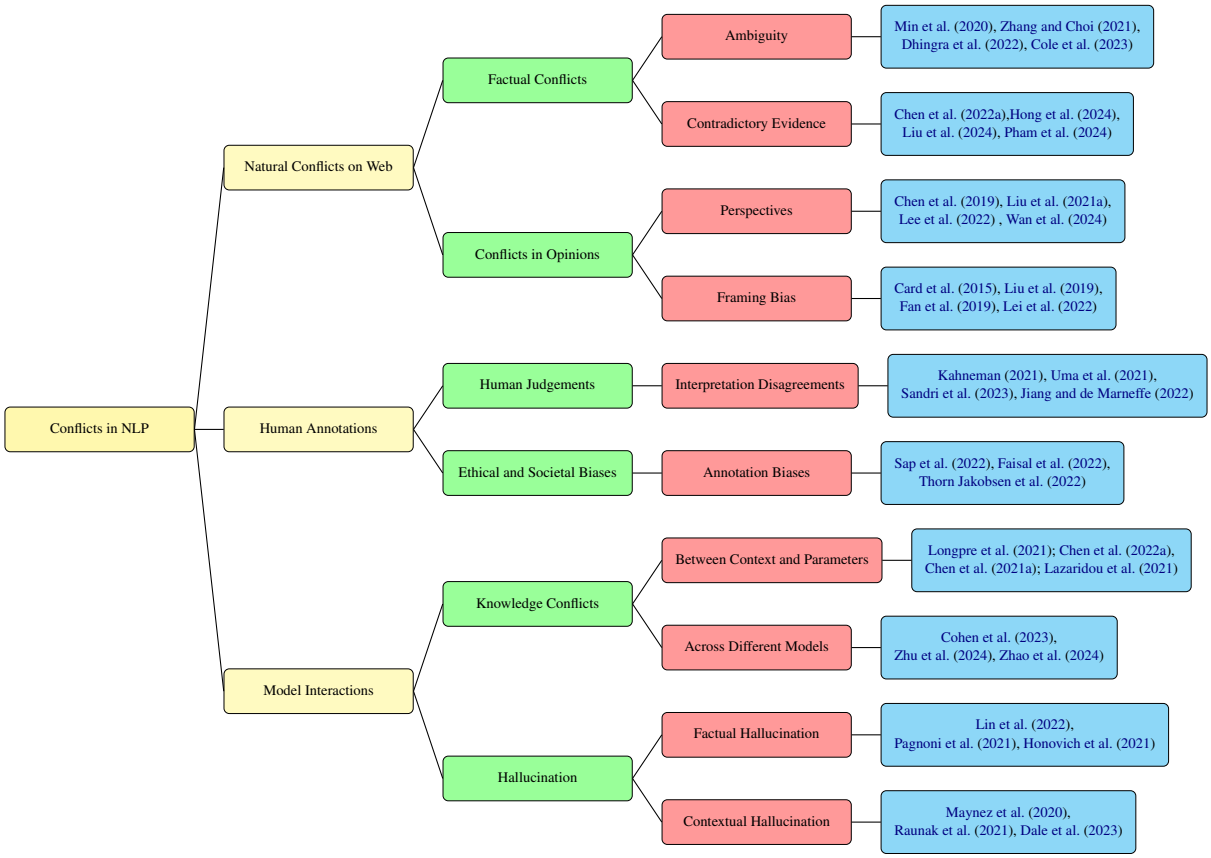

Figure 2: Taxonomy of conflicts in texts.

directions for conflict-aware AI systems.

The abundance of **online data** is accompanied by inherent conflicts, stemming from diverse sources, interpretations, and biases. These conflicts manifest as *factual conflicts*, such as semantic ambiguities (Pavlick and Tetreault, 2016; Min et al., 2020) and factual inconsistencies (Pham et al., 2024; Liu et al., 2024), or as *conflicts in opinions* related to political ideologies (Entman, 1993; Recasens and et al., 2013) and perspectives (Chen et al., 2019; Liu et al., 2021a). Factual conflicts are particularly prevalent in open-domain question answering (QA) and retrieval-augmented generation (RAG) systems (Chen et al., 2017), where aggregating knowledge from multiple sources introduces inconsistencies (Liu et al., 2024). These challenges highlight the need for conflict-aware retrieval and reasoning mechanisms to improve model reliability (Xie et al., 2024). Unlike factual conflicts, opinionated disagreements reflect the variability in human interpretation, beliefs, and ideological stances (Chen et al., 2019; Fan et al., 2019). The presence of conflicting viewpoints complicates tasks such as summarization, sentiment analysis, and dialogue generation, where maintaining coherence and neutrality is crucial (Liu et al., 2021a; Lee et al., 2022). Furthermore, the uneven distribution and biases of web data also affects models to behave from a Western perspective (Ramaswamy et al., 2023; Mihalcea et al., 2024).

Another significant conflict arises in **human-annotated data**. For instance, *annotation disagreements* persists in both subjective and seemingly objective NLP tasks (Mostafazadeh Davani et al., 2022). Disagreements are widespread in sentiment analysis (Wan et al., 2023), hate speech detection (Sap et al., 2022), and even natural language inference (NLI) (Pavlick and Kwiatkowski, 2019). Models trained on aggregated (e.g. majority-vote) labels struggle with ambiguous or high-disagreement examples, often treating them as hard-to-learn or mislabeled (Anand et al., 2023). Pavlick and Kwiatkowski (2019) also find that standard NLI models' uncertainty does not reflect the true ambiguity present in human opinions, leading to overconfidence in contentious cases. In addition, *annotation biases*—such as those related to race, gender, and geography—skew model predictions and reinforce societal biases (Buolamwini and Gebru, 2018; Sap et al., 2022; Pei and Jurgens, 2023).

These issues highlight the need for fair and representative annotations that capture the complexity of human disagreement.

Conflicts also emerge during **interactions with models**, manifesting as *knowledge conflicts* between model memories and contexts, and *hallucinations* in generated outputs. Knowledge conflicts arise when a model's internal memory contradicts external contextual evidence, as shown by Longpre et al. (2021), who found that models often overly depend on memorized knowledge, leading to hallucinations. Neeman et al. (2023) proposed separating parametric and contextual knowledge to improve interpretability, while Xie et al. (2024) examined LLMs' confirmation bias, showing how models inconsistently handle contradictory evidence. Additionally, hallucinations—ranging from factual inconsistencies (Lin et al., 2022; Ouyang and et al., 2022) to contextual hallucinations (Maynez et al., 2020; Kryscinski et al., 2020)—further undermine model reliability. Various mitigation strategies have been proposed, including retrieval augmentation (Lewis et al., 2020; Shuster et al., 2021), hallucination detection (Manakul et al., 2023), and knowledge graph-based verification (Guan et al., 2024).

By systematically categorizing and analyzing these conflicts, this survey provides a unified perspective on their origins, implications, and mitigation strategies. Addressing these issues is essential for building robust and trustworthy AI systems that operate effectively across diverse domains and user groups. Our findings contribute to the development of conflict-aware frameworks for data collection, model training, and model usage, ultimately enhancing the fairness and reliability of NLP.

2 Conflicts in Natural Texts on the Web

Conflicts in natural texts on the web manifest in diverse ways, reflecting the inherent complexity and subjectivity of human language. They can broadly be categorized into factual conflicts, which revolve around factual discrepancies caused by various reasons, and conflicts in opinions, which pertain to divergent perspectives or biases.

2.1 Factual Conflicts

2.1.1 Origins

Ambiguity Ambiguity is a root cause of factual conflict. When a query or piece of data lacks clarity about entities or context, a model can produce conflicting answers. A clear demonstration of how ambiguity induces conflicts is context dependence. For example, an ambiguous question of "which COVID-19 vaccine was the first to be authorized by our government?" can have conflicting answers depending on different geographical contexts (Zhang and Choi, 2021).

Min et al. (2020) was the first work to study the effects of ambiguity in open domain question answering. They introduced AmbigQA, a dataset highlighting that over half of the open-domain, natural questions are ambiguous, with diverse sources of ambiguity such as event and entity references. Zhang and Choi (2021) proposed the SituatedQA task, showing that a significant fraction of open-domain questions are valid only under particular temporal or geographic contexts. Many other work specifically focus on the temporal aspect of ambiguity, benchmarking and evaluating models' awareness and adaptation to time-sensitive questions (Chen et al., 2021b; Liska et al., 2022; Kasai et al., 2023).

Contradictory Evidence Conflicts in NLP systems arise when information on the web presents conflicting evidence towards a factual question. This issue is particularly prevalent in open-domain question answering settings, where models must navigate inconsistencies across diverse information sources. Liu et al. (2024) find that 25% of unambiguous factual questions queried on Google retrieve conflicting evidence from multiple sources. For instance, a Google search for "When was Kendrick Lamar's first album released?" yields conflicting evidence, illustrating the challenge of integrating contradictory information in retrieval-based QA systems (Liu et al., 2024).

Researchers have proposed different datasets to systematically study how NLP models handle such conflicts. Li et al. (2024b) introduce ContraDoc, a human-annotated dataset of long documents with internal contradictions; Pham et al. (2024) propose WhoQA, a benchmark dataset that constructs conflicts by formulating questions about a shared property among entities with the same name (e.g. "Who is George Washington?"); and Liu et al. (2024) construct QACC, a human-annotated dataset of conflicting results retrieved by Google. Beyond empirical datasets, several studies have proposed synthetic approaches to simulate conflicts through entity substitution (Chen et al., 2022a; Hong et al., 2024), machine-generated conflicting

evidence (Pan et al., 2023; Wan et al., 2024; Hong et al., 2024), and pre-defined rule-based templates (Kazemi et al., 2023).

2.1.2 Implications and Mitigation

Zhang and Choi (2021) show that pre-trained language models perform competitively at identifying whether a question is context dependent, matching human agreements, but lag behind human-level performance by a significant margin when answering questions dependent on temporal contexts. Cole et al. (2023) demonstrate that large language models benefit from a "disambiguate-then-answer" pipeline, showing improved reliability by detecting ambiguity before attempting an answer. Dhingra et al. (2022) propose time-aware language models that condition on timestamps to mitigate confusion arising from outdated facts or evolving knowledge.

Studies have shown that retrieval performance heavily impacts which sources models rely on (Chen et al., 2022a), knowledge conflicts within contexts significantly degrade LLMs' performance in RAG settings (Pham et al., 2024; Liu et al., 2024; Li et al., 2024b), and QA models are vulnerable to even small amounts of evidence contamination brought by misinformation (Pan et al., 2023). In addition, large language models (LLMs) exhibit a strong confirmation bias, favoring external evidence that aligns with their parametric memory, even when conflicting evidence is present (Xie et al., 2024). To mitigate these issues, Hong et al. (2024) proposed fine-tuning a discriminator or prompting GPT-3.5 to elicit its discriminative capability and show that these approaches significantly enhance model robustness. On the other hand, Liu et al. (2024) proposed fine-tuning LLMs with human-written explanations to teach models to reason through conflicting evidence.

2.2 Conflicts in Opinions

2.2.1 Origins

Perspectives Individuals and communities often hold diverse perspectives on the same issue. Such diversity is evident in online discussions and debates, where the multiplicity of viewpoints can lead to conflicting opinions. For instance, on controversial topics such as "Animals should have lawful rights," people express varying stances (Chen et al., 2019), posing challenges for downstream tasks like summarization where consolidating viewpoints and presenting unbiased information are crucial (Liu et al., 2021a; Lee et al., 2022).

Chen et al. (2019) address this challenge by presenting a range of perspectives on a given claim. The authors introduce the task of substantiated perspective discovery, where a system identifies diverse, evidence-supported perspectives that take a stance on a claim, and curated a dataset, PERSPECTRUM, for the task, using online debate platforms and search engines. Wan et al. (2024) introduce ConflictingQA, a dataset comprising controversial questions paired with real-world evidence documents presenting divergent facts, argumentation styles, and conclusions. Plepi et al. (2024) study perspective-taking in contentious online conversations (e.g. social media dilemmas). The authors create a new corpus of 95k conflict scenarios augmented with each user's self-disclosed background information. Liu et al. (2021a) propose MultiOpEd, an open-domain corpus focusing on automatic perspective discovery. MultiOpEd comprises 1,397 controversial topics, each accompanied by two editorials expressing opposing viewpoints, and each editorial includes a one-sentence perspective summarizing its core argument and a brief abstract highlighting supporting details.

Framing Bias A specific example of how differing opinions are conveyed and expanded is framing bias, a mechanism in which news media shape interpretations by emphasizing certain aspects of information over others (Entman, 1993). In a polarized media environment, partisan media outlets deliberately frame news stories in a way to advance certain political ideologies (Jamieson et al., 2007; Levendusky, 2013; Liu et al., 2019).

Various studies have examined different facets of such media bias. Card et al. (2015) introduce the Media Frames Corpus (MFC) to facilitate the computational study of media framing in news articles. The MFC comprises several thousand news articles covering three policy issues, each annotated according to 15 general-purpose framing dimensions. Liu et al. (2019) introduce Gun Violence Frame Corpus (GVFC), a dataset of news headlines with frames curated and annotated by journalism and communication experts. Fan et al. (2019) focus on informational bias, where factual content is subtly framed by content selection and organization within articles. They introduce BASIL, a dataset comprising 300 news articles annotated with 1,727 bias spans, revealing that informational bias occurs more frequently than lexical bias. Lei et al. (2022) identify ideological bias at the sentence level by

analyzing news discourse structure, showing that bias-indicative sentences may appear neutral in isolation.

2.2.2 Implications and Mitigation

Analysis of PERSPECTRUM reveals significant natural language understanding challenges, as human performance substantially outperforms machine baselines at identifying diverse, evidence-supported perspectives (Chen et al., 2019). Furthermore, when selecting real-world evidence for controversial questions, LLMs predominantly prioritize the relevance of the evidence to the query, often disregarding stylistic attributes such as the presence of scientific references or a neutral tone (Wan et al., 2024). In addition, the distribution and biases of web data also affects models to behave from a Western perspective (Ramaswamy et al., 2023; Mihalcea et al., 2024). Studies have shown that LLMs' outputs skew toward the values of Western English-speaking countries (Tao et al., 2024; Naous et al., 2024), and misalignment is more pronounced for underrepresented personas and on culturally sensitive topics such as social values (Al Kuwatly et al., 2020). Furthermore, LLMs often provide inconsistent answers to the same question when prompted in different languages (Li et al., 2024a; AlKhamissi et al., 2024), revealing conflicting cultural perspectives within a single model.

Several studies have proposed methods to address conflicts in perspectives. Liu et al. (2021a) show that incorporating auxiliary tasks enhances the quality of perspective summarization. Chen et al. (2022b) propose a novel document retrieval paradigm that focuses aggregating and displaying responses from web documents based on varying viewpoints, and reveal that users prefer seeing search results in different clusters of perspectives instead of in a list ranked by relevance. Jiang et al. (2023) tackle opinion summarization for user reviews by generating summaries from different perspectives. Their framework selects subsets of reviews based on sentiment polarity and informational contrast, and produces balanced pros, cons, and verdict summaries. Plepi et al. (2024) demonstrate that a tailored generation model which conditions on a user's personal context produces more appropriate and empathetic responses than large general models, effectively capturing different viewpoints in the conflict.

To mitigate framing bias and ideology conflicts, Milbauer et al. (2021) propose a method to uncover complex ideological and worldview differences across online communities beyond a single left-vs-right axis. Their study identifies multiple axes of polarization and nuanced ideological distinctions, offering a more multifaceted analysis of online opinion differences. Liu et al. (2022) introduce a pre-training approach for ideology classification that learns ideological bias by directly comparing news articles about the same event reported by outlets with different leanings. Chen et al. (2023) tackle political ideology classification under limited and biased data by disentangling content from style, enabling accurate ideology detection with minimal training data. Lee et al. (2022) present a model that leverages news titles and employs hierarchical multi-task learning to neutralize biased content from title to article, while Liu et al. (2023) induce a neutral event graph that captures events with minimal framing bias by synthesizing across different ideologies.

3 Conflicts in Human-Annotated Texts

3.1 Origins

Annotation Disagreement The subjective nature of human judgments bring noise and disagreements to their annotated data (Kahneman, 2021). Such annotation disagreements in NLP arise from multiple sources, including linguistic ambiguity, annotator backgrounds, task design, and dataset curation practices. Uma et al. (2021) provide a comprehensive survey of disagreement across NLP and computer vision tasks, highlighting subjective ambiguity and annotator diversity as key factors of disagreements. Sandri et al. (2023) categorize disagreements in offensive language detection, showing that some stem from inherent ambiguity while others result from annotation errors or lack of context. Their findings imply that not all disagreements are equal – some signal hard-to-classify content, whereas others indicate correctable annotation issues. Jiang and de Marneffe (2022) similarly classify NLI disagreements into three broad categories—linguistic uncertainty, annotator bias, and task design issues. They show that both linguistic uncertainty and annotator bias contribute substantially to label variability, and a significant portion of "disagreement noise" in NLI is systematic and predictable (e.g. stemming from specific ambiguity types or annotator profiles).

Task design also significantly influences annotation disagreements. Dsouza and Kovatchev (2025)

find that labels disagreement in Reinforcement Learning from Human Feedback (RLHF) largely depend on annotator selection and task formulation, while Yung and Demberg (2025) show that free-choice annotation methods, where annotators select any suitable connective, yield less diversity (often converging on common labels) than forced-choice approaches, where annotators choose from predefined options, in discourse relation labeling.

On the other hand, demographic and ideological factors also shape disagreement. Pavlick and Kwiatkowski (2019) suggest that many NLI disagreements are not errors but reflect genuine ambiguities in language and individual variations in background knowledge. Sap et al. (2022) demonstrate that annotators' personal beliefs and identities affect their perception of toxicity, and Wan et al. (2023) find that demographic data significantly enhances the prediction of annotation disagreements.

Ethical and Societal Biases Ethical and societal biases are also present in human-annotated texts. Biases introduced during annotation, whether related to race, gender, or geography, can significantly skew model predictions and decision-making processes (Buolamwini and Gebru, 2018). For instance, studies have shown that popular NLP datasets have a severe Western-centric skew (Faisal et al., 2022). This Western dominance in training/evaluation data means models optimized on these benchmarks assume a Western context by default. A model might perform well on answering questions about New York or London (since those appear often in the data), but fail on questions about Nairobi or Manila simply due to lack of exposure (Faisal et al., 2022).

Sap et al. (2022) show that annotators' ideological and racial attitudes affect their judgments of toxicity, with conservative annotators less likely to flag anti-Black slurs as toxic but more likely to misclassify African American English (AAE) as offensive. Similarly, Thorn Jakobsen et al. (2022) examine how annotation guidelines interact with annotator demographics, finding that different task formulations can either exacerbate or mitigate biases. They show that even well-designed guidelines can elicit systematically different responses from distinct demographic groups, underscoring the importance of inclusive task framing to reduce disparities in annotations. Pei and Jurgens (2023), on the other hand, introduce POPQUORN, a dataset explicitly designed to measure the impact of annotator demographics across multiple NLP tasks. Their large-scale analysis confirms that annotator background, such as age, gender, race, and education, accounts for substantial variance in annotations.

3.1.1 Implications and Mitigation

Early research has highlighted the impact of annotator disagreements on data quality and model performance (Artstein and Poesio, 2008; Pustejovsky and Stubbs, 2012; Plank et al., 2014). Pavlick and Kwiatkowski (2019) find that standard NLI models' uncertainty did not reflect the true uncertainty found in human opinions. This mismatch suggests that when disagreements are ignored, models become overconfident on contentious cases. In addition, Anand et al. (2023) observe that a classifier provided only with single "gold" labels tends to be less confident and less accurate on examples where annotators widely disagreed on. Such models may treat these instances as "hard to learn" or mislabeled. Furthermore, Sap et al. (2019) reveal how annotator biases can lead to racist outcomes in hate speech classifiers. They uncover a spurious correlation: tweets written in African American English (AAE) are often rated as more toxic by annotators, even when they are not hate speech. Models trained on such data inherit this bias, falsely flagging content by Black authors as offensive at disproportionately high rates.

In terms of data collection, previous work has explored the strategy of acquiring multiple labels for each data item from various annotators to enhance data quality. They develop a probabilistic model to assess the true label of an item by considering the varying expertise of annotators and the possibility of label noise (Sheng et al., 2008). Mostafazadeh Davani et al. (2022) examine strategies to train NLP models without discarding annotator disagreements. They propose a multi-annotator modeling approach: a multi-task neural network that learns to predict each individual annotator's label as a separate output while sharing a common representation. Similarly, different studies have shown that models which incorporate annotator disagreement as "soft" labels (i.e. full label distribution) for training outperform those trained on aggregated single labels (Uma et al., 2021; Fornaciari et al., 2021)

4 Conflicts during Model Interactions

4.1 Knowledge Conflicts

4.1.1 Origins

Context vs. Memory A common type of knowledge conflict arises when a model's prompt (contextual knowledge) contradicts what the model has learned and stored in its parameters (parametric knowledge) (Longpre et al., 2021; Chen et al., 2022a). One prevalent cause of such conflicts is the presence of updated information (Chen et al., 2021a; Lazaridou et al., 2021; Luu et al., 2022), where newly available knowledge contradicts models' previously learned knowledge.

Recent studies have developed many evaluation frameworks and datasets to assess LLMs' behaviors in this scenario through different methods, including entity substitution (Longpre et al., 2021; Chen et al., 2022a; Wang et al., 2024), adversarial perturbation (Chen et al., 2022a; Xie et al., 2024), misinformation injection (Pan et al., 2023), and machine generation (Qian et al., 2023; Ying et al., 2024; Tan et al., 2024).

Within and Across Models Different models may exhibit conflicts in their knowledge bases. Cohen et al. (2023) investigate how different LLMs can be used to fact-check each other to reveal inconsistencies that imply factually incorrect claims. On the other hand, Zhu et al. (2024) address the issue of inconsistencies between the visual and language components in Large Vision-Language Models (LVLMs). These conflicts arise due to the separate training processes and distinct datasets used for each modality, leading to discrepancies in the knowledge they capture. Even the same model may have internal knowledge conflicts. Zhao et al. (2024) identify intra-model contradictions by paraphrasing the same query multiple times and find answer divergence across different LLMs.

4.1.2 Implications and Mitigation

Interestingly, different studies of knowledge conflicts present seemingly contradictory findings. Some studies claim that models often excessively rely on parametric memory when observing conflicts with contextual knowledge (Longpre et al., 2021); Some other studies posit that LLMs tend to ground their answers in retrieved documents in this scenario (Chen et al., 2022a; Qian et al., 2023; Tan et al., 2024); or even both – LLMs are highly receptive to context when it is the only evidence presented in a coherent way, but also demonstrate a strong confirmation bias toward parametric memory when both supportive and contradictory evidence to their parametric memory are present (Xie et al., 2024).

Various approaches have been proposed to mitigate the consequences of such knowledge conflicts. Longpre et al. (2021) propose a simple method that augments the training set with training examples modified by corpus substitution to mitigate memorization, Chen et al. (2022a) present a calibrator that abstains from predicting on instances with conflicting evidence, and Wang et al. (2024) propose a new instruction-based approach that augment LLMs to first identify knowledge conflicts, then pinpoint conflicting information segments, and lastly provide distinct answers in conflicting scenarios.

4.2 Hallucination

4.2.1 Origins

Factual Hallucinations Factual hallucinations occur when a model's output conflicts with real-world facts. TruthfulQA (Lin et al., 2022) introduces an adversarial question-answering benchmark, revealing that even the best model at that time (GPT-3) was truthful on only 58% of questions, compared to 94% for humans. Pagnoni et al. (2021) construct the FRANK dataset, which annotates factual errors in summarization, identifying strengths and weaknesses of various metrics. Similarly, Honovich et al. (2021) extend QAGS to knowledge-grounded dialogue using question generation and entailment modeling to assess factual consistency. To further evaluate LLMs' factual knowledge and reasoning, Hu et al. (2023) introduce Pinocchio, a benchmark comprising 20,713 multiple-choice questions across various domains, timelines, and languages, and assess models on fact composition, temporal reasoning, and adversarial robustness, revealing that LLMs often struggle with factual consistency and are prone to spurious correlations. Mallen et al. (2023) also find that language models struggle with less popular factual knowledge, and that retrieval augmentation helps significantly in these cases.

Contextual Hallucinations Contextual hallucinations occur when generated text contradicts the given input context, such as in summarization, translation, and generation tasks. Maynez et al. (2020) find that summarization models frequently

generate content unfaithful to input documents, with 64% of summaries containing unsupported information. In machine translation, Raunak et al. (2021) analyze hallucinations caused by source perturbations and training noise, and find that slight modifications to input data could trigger off-topic translations. Similarly, Dale et al. (2023) introduce HalOmi, a multilingual benchmark for hallucination and omission detection in machine translation, showing that prior hallucination detectors often fail across different language pairs. In generation tasks, Liu et al. (2021b) propose a novel token-level, reference-free hallucination detection task and dataset (HADES) for free-form text generation, and Niu et al. (2024) introduce RAGTruth, a comprehensive corpus designed for analyzing word-level hallucinations across various domains and tasks within standard Retrieval-Augmented Generation (RAG) frameworks.

4.2.2 Implications and Mitigation

Mitigating hallucinations in language models has been approached through various strategies, including knowledge disentanglement (Neeman et al., 2023), retrieval augmentation (Lewis et al., 2020; Shuster et al., 2021), knowledge graphs (Guan et al., 2024), and improved verification methods (Kryscinski et al., 2020; Wang et al., 2020; Laban et al., 2022; Manakul et al., 2023). DisentQA enhances robustness by training models to separate internal memory from external context, improving accuracy in conflicting knowledge scenarios (Neeman et al., 2023). Retrieval-Augmented Generation (RAG) mitigates factual inconsistencies by integrating external sources like Wikipedia (Lewis et al., 2020) or incorporating a neural search module into chatbot responses (Shuster et al., 2021). In addition, Guan et al. (2024) demonstrate how retrofitting LLM outputs using structured knowledge graphs can correct factual inconsistencies, particularly in complex reasoning tasks. For hallucination detection methods, FactCC and QAGS introduce automated methods using synthetic data and question-answer validation to assess factual consistency (Kryscinski et al., 2020; Wang et al., 2020). SummaC refines entailment-based scoring (Laban et al., 2022), and SelfCheckGPT detects hallucinations by sampling multiple model outputs and checking for agreement without external references (Manakul et al., 2023).

5 Open Challenges

Building Conflict-Aware AI Systems Conflicts in training data and retrieved contexts can lead to unexpected consequences in NLP models, affecting their reliability and trustworthiness (Pan et al., 2023; Liu et al., 2024). Studies show that knowledge conflicts within retrieved contexts severely degrade LLM performance in retrieval-augmented generation (RAG) (Chen et al., 2022a; Pham et al., 2024; Liu et al., 2024; Li et al., 2024b), and QA models are highly susceptible to misinformation, where even minimal evidence contamination can lead to incorrect predictions (Pan et al., 2023). Also, LLMs exhibit confirmation bias, where models prefer external evidence that aligns with their parametric knowledge, even when conflicting information is present (Xie et al., 2024). This bias can cause models to reinforce incorrect or outdated information instead of updating their knowledge based on newly retrieved sources. We highlight the need of developing conflict-aware NLP systems that (1) assess and resolve contradictions in retrieved contexts, (2) mitigate confirmation bias by designing models that reason over conflicting evidence, and (3) integrate fact-checking mechanisms that evaluate source credibility and detect misinformation before generating responses.

Towards Robust and Fair NLP Models Beyond technical challenges in handling knowledge conflicts, systemic biases in training data and annotation processes impact the fairness of NLP models. The distribution of web data skews LLM outputs toward Western perspectives, disproportionately reflecting Western English-speaking values (Ramaswamy et al., 2023; Mihalcea et al., 2024; Tao et al., 2024; Naous et al., 2024). Furthermore, LLMs provide inconsistent answers across languages, revealing contradictions in how they encode cultural perspectives (Li et al., 2024a; AlKhamissi et al., 2024). Biases also emerge from annotation processes, where demographic and ideological factors influence how data is labeled (Sap et al., 2019). Future directions to create robust and fair NLP models include (1) building demographically diverse and geographically representative training datasets, (2) enhancing consistency of models by aligning cross-lingual knowledge representations, and (3) improving annotation frameworks to account for demographic and ideological biases among annotators.

Limitations

Conflicting information is present both in the data that models rely on and in their generated outputs. While we strive to account for all potential conflict scenarios, some cases may inevitably be overlooked. Additionally, due to space constraints, we do not provide an exhaustive discussion of the literature on each specific type of conflict. Instead, we adopt a broader perspective, examining various types of conflicts to identify connections, patterns, and common challenges.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Abhishek Anand, Negar Mokhberian, Prathyusha N. Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2023. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. *arXiv preprint arXiv:2403.04085*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Chen Chen, Dylan Walker, and Venkatesh Saligrama. 2023. Ideology prediction from scarce and biased supervision: Learn to disregard the "what" and focus on the "how"! In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9529–9549, Toronto, Canada. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022a. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2022b. Design challenges for a multi-perspective search engine. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 293–303, Seattle, United States. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021a. A dataset for answering time-sensitive questions.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021b. A dataset for answering time-sensitive questions.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination.

Robert Cole, Alice Wu, and Scott Frazier. 2023. Selective question answering under ambiguity: Improving llm reliability by disambiguating then answering. *arXiv preprint arXiv:2308.01234*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and

William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Russel Dsouza and Venelin Kovatchev. 2025. Sources of disagreement in data for LLM instruction tuning. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 20–32, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43:51–58.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. 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 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting liu. 2024. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts?

Kathleen Hall Jamieson, Bruce W Hardy, and Daniel Romer. 2007. The effectiveness of the press in serving the needs of american democracy. *Institutions of American democracy: A republic divided*, (717):21–51.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023. Large-scale and multi-perspective opinion summarization with diverse review subsets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5641–5656, Singapore. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

D Kahneman. 2021. *Noise: a flaw in human judgment*. HarperCollins.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Philippe Laban, Joey Tiao, Arman Cohan, and Iz Beltagy. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew S Levendusky. 2013. Why do partisan media polarize viewers? *American journal of political science*, 57(3):611–623.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024a. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. 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 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024b. ContraDoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021a. MultiOpEd: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361, Online. Association for Computational Linguistics.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2024. Open domain question answering with conflicting contexts.

Siyi Liu, Hongming Zhang, Hongwei Wang, Kaiqiang Song, Dan Roth, and Dong Yu. 2023. Open-domain event graph induction for mitigating framing bias. *arXiv preprint arXiv:2305.12835*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021b. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7052–7063.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2024. Why ai is weird and should not be this way: Towards ai for everyone, with everyone, by everyone. *arXiv preprint arXiv:2410.16315*.

Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4832–4845, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10056–10070.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models.

Long Ouyang and et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 32nd International Conference on Computational Linguistics and 12th International Joint Conference on Natural Language Processing (COLING/IJCNLP)*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Ellie Pavlick and Joel Tetreault. 2016. Semantically motivated future directions in linguistic ambiguity detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice.

Barbara Plank, Dirk Hovy, Anders Sogaard, et al. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of EACL*. Association for Computational Linguistics.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective taking through generating responses to conflict situations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6482–6497, Bangkok, Thailand. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Marta Recasens and et al. 2013. Linguistic models for analyzing and detecting bias. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of NAACL-HLT 2022*, pages 5884–5906. Association for Computational Linguistics. Shows strong correlations between annotators' demographic/political identities and their toxicity annotations, highlighting the need to account for annotator perspectives8203;:contentReference[oaicite:108]index=108.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Terne Sasha Thorn Jakobsen, Maria Barrett, Anders Søgaard, and David Lassen. 2022. The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 44–61, Marseille, France. European Language Resources Association.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *arXiv preprint arXiv:2402.11782*.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13718–13726. Demonstrates that incorporating annotators' demographic background features helps predict and explain label disagreements in multiple subjective NLP datasets8203;:contentReference[oaicite:109]index=109.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Resolving knowledge conflicts in large language models.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating LLMs' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246, Bangkok, Thailand. Association for Computational Linguistics.

Frances Yung and Vera Demberg. 2025. On crowdsourcing task design for discourse relation annotation. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing what llms do not know: A simple yet effective self-detection method.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models.