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 for developing conflict-aware and robust NLP systems, and propose concrete research directions to address them.

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).

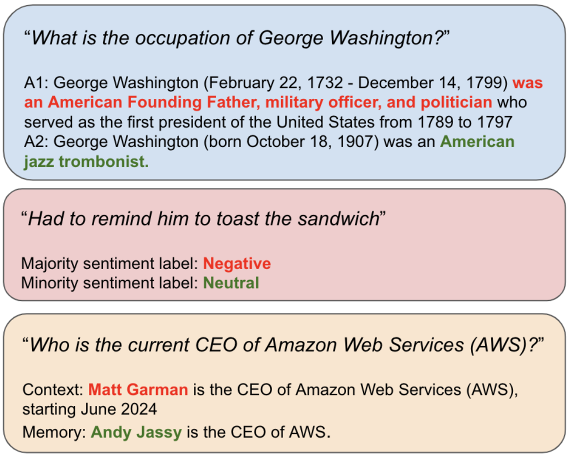

Figure 1: Examples of the three different areas of conflicts discussed in this work. The first example describes a case where two different entities of the same name are found *naturally on the web*, the second example elaborates the *annotation disagreement* in a sentiment analysis task, and the third 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; Klie et al., 2023), 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.

To ensure comprehensive and representative coverage of conflicts in NLP, we first establish a high-level categorization encompassing three primary sources: (1) natural conflicts present in web data, (2) conflicts arising from human annotation,

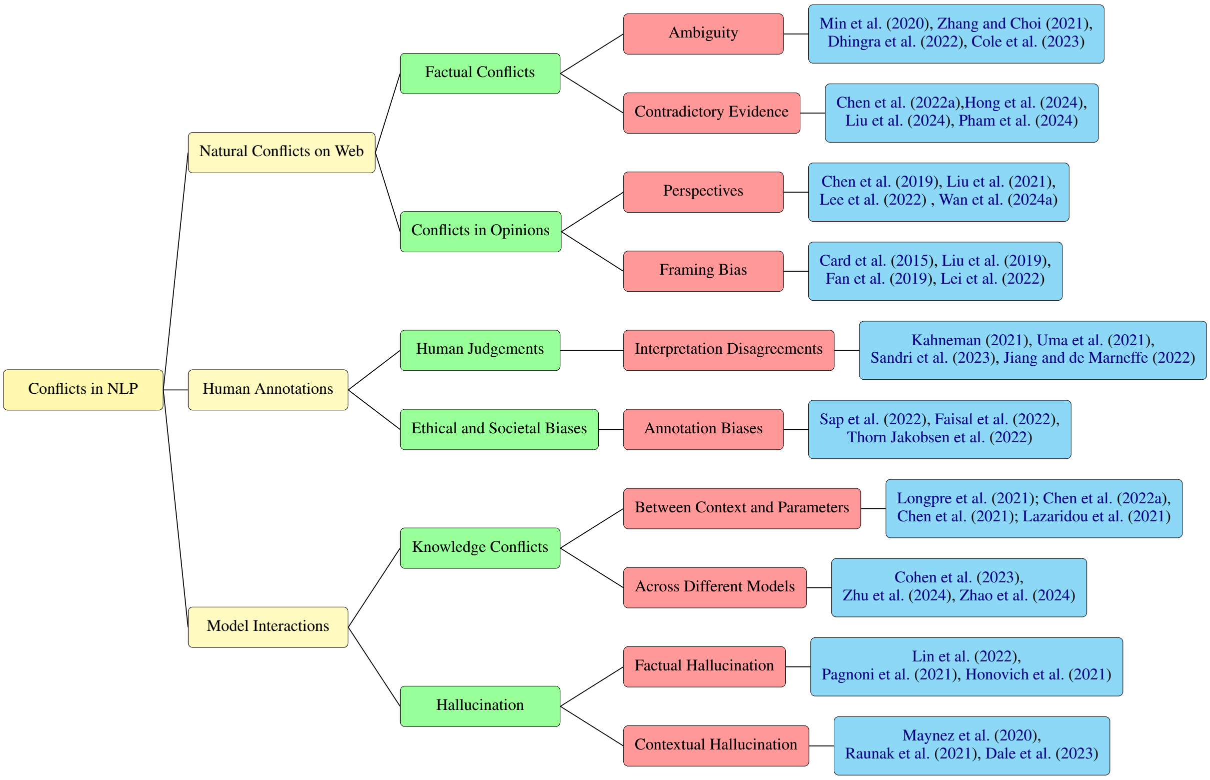

Figure 2: Taxonomy of conflicts in texts.

and (3) conflicts emerging from model interactions. Notably, conflicts found in natural web texts and human-annotated datasets are primarily present in the training data—i.e., the inputs to models—whereas conflicts involving model interactions can arise in various forms, such as inconsistencies between model outputs and their inputs, contradictions among multiple outputs, or conflicts within the outputs themselves. For each category, we identify influential and widely cited survey papers as initial seed works (Uma et al., 2021; Klie et al., 2023; Zhang et al., 2023; Xu et al., 2024; Feng et al., 2024; Wang et al., 2023). Building upon these seeds, we systematically trace and incorporate the most impactful and representative studies for each type of conflict through citation chaining and targeted literature searches across major databases. This approach enables us to synthesize developments in each category and connect them, thereby providing an integrated discussion of current challenges, impacts on downstream tasks, and promising future 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., 2021). 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., 2021; 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., 2025).

Another significant conflict arises in **human-annotated data**. For instance, *annotation dis-*

agreements 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., 2024). 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 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).

In this survey, we systematically examine the landscape of conflicts in NLP by categorizing them into three primary sources. For each conflict type, we detail how such conflicts arise and in what forms they take (**origins**), the challenges they pose (**implications**), and the strategies developed to address them (**mitigation**). We present a comprehensive taxonomy in Figure 2, as well as structured summary tables—Table 1, Table 2, and Table 3—that synthesize datasets, methodologies, and analysis from prior work. By offering a unified framework for understanding and addressing conflicting information in NLP, this survey contributes 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., 2021; 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. For example, Liu et al. (2024) find that 25% of unambiguous factual questions queried on Google retrieve conflicting evidence from multiple sources.

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

2.1.2 Implications and Mitigation

Implications Factual conflicts pose significant challenges for NLP systems. Pre-trained language models accurately detect context-dependent questions but fall short when answering queries requiring temporal context, performing notably below human levels (Zhang and Choi, 2021). Additionally, large language models (LLMs) often exhibit confirmation bias, favoring retrieved information that aligns with their parametric memory despite contradictory evidence (Xie et al., 2024). Consequently, conflicting information sources severely impact retrieval-augmented generation (RAG) frameworks, significantly degrading model performance even with minimal misinformation exposure (Pham et al., 2024; Liu et al., 2024; Li et al., 2024b; Pan et al., 2023).

Mitigation To address these challenges, various mitigation strategies have been proposed. Effective methods include fine-tuning calibrators for selective abstention (Chen et al., 2022a), employing a "disambiguate-then-answer" pipeline to detect ambiguity proactively (Cole et al., 2023), and developing time-aware models that condition responses on timestamps to manage outdated information (Dhingra et al., 2022). Further robustness improvements have been achieved through fine-tuning discriminators or prompting GPT-3.5 models to explicitly recognize conflicting evidence (Hong et al., 2024), as well as incorporating human-written explanations in fine-tuning processes to enhance models' reasoning capabilities (Liu et al., 2024).

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., 2021; Lee et al., 2022).

Several studies have explored perspectives in the context of conflicting information. Chen et al. (2019) introduce the task of substantiated perspective discovery, where systems identify diverse, evidence-supported stances on a claim, and release the PERSPECTRUM dataset using online debates and search results. Wan et al. (2024a) propose ConflictingQA, a dataset of controversial questions paired with real-world documents that present divergent facts, arguments, and conclusions. Plepi et al. (2024) examine perspective-taking in contentious online discourse, curating a corpus of 95k conflict scenarios annotated with users' self-reported backgrounds. Liu et al. (2021) present MultiOpEd, a corpus of 1,397 controversial topics, each paired with opposing editorials and concise summaries capturing their core perspectives.

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).

Numerous studies have investigated different aspects of media bias. Card et al. (2015) introduce the Media Frames Corpus (MFC), a collection of news articles annotated with 15 general-purpose framing dimensions across three policy issues, enabling computational analysis of media framing. Liu et al. (2019) present the Gun Violence Frame Corpus (GVFC), a dataset of news headlines annotated by domain experts to capture framing in gun violence reporting. Fan et al. (2019) examine informational bias—bias conveyed through content selection and structure—and release BASIL, a dataset of 300 news articles annotated with 1,727 bias spans, demonstrating that informational bias is more prevalent than lexical bias.

2.2.2 Implications and Mitigation

Implications 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., 2024a). 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., 2025). 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; Eloundou et al., 2025), revealing conflicting cultural perspectives within a single model.

Mitigation Several studies have proposed methods to address conflicts in perspectives and ideological bias. Liu et al. (2021) show that auxiliary tasks improve perspective summarization quality, while Chen et al. (2022b) propose a retrieval paradigm that clusters documents by viewpoint, revealing users' preference for diverse perspectives over relevance-ranked lists. Jiang et al. (2023b) generate opinion summaries by selecting review subsets based on sentiment polarity and contrast, producing balanced pros, cons, and verdicts. Plepi et al. (2024) demonstrate that conditioning generation on users' personal contexts yields more empathetic and appropriate responses than general-purpose models.

To mitigate framing and ideological bias, Milbauer et al. (2021) uncover nuanced worldview differences across communities by identifying multiple axes of polarization beyond the traditional left–right spectrum. Liu et al. (2022b) pre-train models for ideology detection by comparing reporting on the same events across partisan sources. Chen et al. (2023) disentangle content from style to enable ideology classification under data scarcity and bias. Lee et al. (2022) employ hierarchical multi-task learning to neutralize bias from news titles to article bodies, while Liu et al. (2023) construct neutral event graphs by synthesizing perspectives across ideological divides.

3 Conflicts in Human-Annotated Texts

Conflicts in human-annotated texts largely arise from two sources: annotation disagreements and societal or ethical biases. Disagreements stem from linguistic ambiguity, annotator backgrounds, and task design, while biases reflect systematic demographic or ideological influences that can skew labeling in consistent ways. Though conceptually distinct, these sources often interact—biases may amplify disagreement or entrench disparities. Differentiating between them is essential for understanding annotation-related conflicts and for developing more reliable and equitable NLP datasets.

3.1 Origins

Annotation Disagreement The subjective nature of human judgment introduces variability and disagreement into annotated data (Kahneman, 2021). In NLP, such disagreements arise from linguistic ambiguity, annotator backgrounds, task design, and dataset curation practices. Uma et al. (2021) survey disagreements across NLP and vision tasks, identifying subjective ambiguity and annotator diversity as key contributors. Sandri et al. (2023) classify disagreements in offensive language detection as stemming from inherent ambiguity, annotation errors, or contextual gaps, highlighting that some disagreements reflect hard-to-classify content, while others indicate correctable issues. Similarly, Jiang and de Marneffe (2022) categorize NLI disagreements into linguistic uncertainty, annotator bias, and task design, showing that much of the observed noise is systematic and predictable.

Task formulation also plays a critical role. Dsouza and Kovatchev (2025) find that label disagreement in reinforcement learning from human feedback (RLHF) is shaped by annotator selection and task phrasing. Demographic and ideological factors further influence disagreements. Pavlick and Kwiatkowski (2019) argue that many NLI disagreements reflect genuine linguistic ambiguity and individual variation rather than annotation error. Sap et al. (2022) demonstrate that annotators' personal beliefs and identities affect toxicity judgments, while Wan et al. (2023) show that demographic features significantly improve disagreement prediction.

Ethical and Societal Biases Human-annotated texts also encode societal biases related to race, gender, and geography, which can significantly skew model predictions and downstream decisions (Buolamwini and Gebru, 2018). Sap et al. (2022) show that annotators' ideological and racial identities influence toxicity judgments, with conservative annotators less likely to flag anti-Black slurs and more likely to misclassify African American English (AAE) as offensive. Thorn Jakobsen et al. (2022) examine how annotation guidelines interact with annotator demographics, demonstrating that even well-designed tasks can elicit systematically different responses across groups, highlighting the need for inclusive task design. Pei and Jurgens (2023) introduce POPQUORN, a dataset designed to assess demographic effects on annotation across NLP tasks, and find that annotator attributes—such as age, gender, race, and education—account for substantial variance in labeling behavior.

3.1.1 Implications and Mitigation

Implications Early research has underscored the impact of annotator disagreement on data quality and model performance (Artstein and Poesio, 2008; Pustejovsky and Stubbs, 2012; Plank et al., 2014). Pavlick and Kwiatkowski (2019) show that standard NLI models fail to capture the true uncertainty present in human judgments, leading to overconfidence on contentious examples. Similarly, Anand et al. (2024) find that models trained on single "gold" labels perform poorly and exhibit lower confidence on high-disagreement instances, often treating them as mislabeled or hard to learn. Sap et al. (2019) demonstrate how annotator bias can yield discriminatory outcomes: tweets in African American English (AAE) are frequently misclassified as toxic, a bias inherited by models that disproportionately flag content from Black authors. Additionally, many widely used NLP datasets exhibit a strong Western-centric skew (Faisal et al., 2022), causing models to generalize poorly to underrepresented regions—for example, excelling on questions about New York or London, but failing on Nairobi or Manila due to lack of exposure.

Mitigation Prior work has explored collecting multiple labels per data item to capture annotation variability and improve data quality. Probabilistic models have been developed to infer true labels by accounting for annotator expertise and label noise (Sheng et al., 2008). Mostafazadeh Davani et al. (2022) propose a multi-task neural network that models each annotator's labels individually while sharing a common representation, preserving disagreement in training. Similarly, studies show that models trained on soft labels—i.e., full label distributions reflecting annotator disagreement—consistently outperform those trained on aggregated labels (Uma et al., 2021; Fornaciari et al., 2021).

4 Conflicts during Model Interactions

Conflicts during model interactions primarily manifest as knowledge conflicts and hallucinations, each posing distinct challenges. Knowledge conflicts occur when a model's parametric memory contradicts contextual input or when inconsistencies arise across models, whereas hallucinations occur when outputs deviate from real-world facts or the given input. Differentiating these two types of conflict clarifies their underlying causes and helps guide targeted mitigation strategies.

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., 2021; 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., 2024; Ying et al., 2024; Tan et al., 2024).

Within and Across Models Conflicts may also arise across or within model knowledge bases. Cohen et al. (2023) explore how different LLMs encode different knowledge and can be used to fact-check one another, uncovering inconsistencies indicative of factual errors. Zhu et al. (2024) examine cross-modality conflicts in vision-language models, attributing discrepancies between visual and textual components to separate training regimes and distinct data sources. Even within a single model, contradictions can emerge: Zhao et al. (2024) detect intra-model inconsistencies by paraphrasing queries and observing divergent answers across prompts.

4.1.2 Implications and Mitigation

Implications 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., 2024; 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).

Mitigation Several approaches have been proposed to mitigate the impact of knowledge conflicts. Longpre et al. (2021) reduce memorization by augmenting training data through corpus substitution. Chen et al. (2022a) introduce a calibrator that abstains from prediction when conflicting evidence is detected. More recently, Wang et al. (2024) propose an instruction-based framework that enables LLMs to identify conflicts, localize conflicting segments, and generate distinct responses for conflicting scenarios.

4.2 Hallucination

4.2.1 Origins

Factual Hallucinations Factual hallucinations arise when a model's output contradicts real-world facts. Lin et al. (2022) present TruthfulQA, an adversarial QA benchmark, and show that even top-performing models like GPT-3 were truthful on only 58% of questions, compared to 94% for humans. Pagnoni et al. (2021) construct FRANK, a dataset for identifying factual errors in summarization, while Honovich et al. (2021) extend QAGS to dialogue by leveraging question generation and entailment for factual consistency evaluation. To assess factual knowledge and reasoning in LLMs, Hu et al. (2024) introduce Pinocchio, a large benchmark covering multiple domains, timelines, and languages, revealing challenges in composition, temporal reasoning, and robustness. Mallen et al. (2023) further find that models struggle with less common factual knowledge, with retrieval augmentation significantly improving performance in such 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. (2022a) 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

Implications Hallucinations threaten trust, safety, and the integrity of AI-powered workflows. Hallucinated outputs can rapidly spread false information. For instance, in 2023, an AI-generated image purporting to show an explosion near the Pentagon went viral, briefly causing public panic and even a stock market dip before being debunked

(Sun et al., 2024). Hallucinations directly degrade the performance of downstream applications like abstractive summarization. Studies have found that a large portion of generated summaries contain unsupported facts, misleading readers and propagating misinformation in news and scientific dissemination (Kryscinski et al., 2020).

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 Connections, Challenges and Directions

Given the significance and impact of conflicts in NLP, we advocate for increased attention to the development of conflict-aware and robust AI systems. In this section, we highlight specific challenges by connecting different types of conflicts and propose concrete research directions to address them.

Culturally Robust LLMs Among the challenges outlined in this survey, the development of culturally robust LLMs remains particularly underexplored. Cultural conflicts emerge both in naturally occurring web data and human-annotated datasets, where Western-centric distributions dominate. Prior studies reveal that LLMs often reflect the values and perspectives of Western, English-speaking populations (Ramaswamy et al., 2023; Mihalcea et al., 2025; Tao et al., 2024; Naous et al., 2024), with misalignments especially pronounced for underrepresented personas and culturally sensitive topics (Al Kuwatly et al., 2020). Additionally, LLMs exhibit inconsistent behavior across languages (Li et al., 2024a; AlKhamissi et al., 2024; Eloundou et al., 2025), revealing internal cultural conflicts. These issues are rooted in the data: both the pre-train data and benchmark datasets commonly exhibit Western-centric biases (Mihalcea et al., 2025; Faisal et al., 2022), causing models to default to Western contexts and perform poorly on less-represented regions and cultures.

However, to the best of our knowledge, no effective methodology has yet been proposed to address this issue. With the emergence of culturally distinct LLMs—such as Qwen, trained largely on Chinese data (Bai et al., 2023), and Vikhr, trained on Russian data (Nikolich et al., 2025)—a promising direction is model fusion across culturally diverse models to achieve greater cultural balance (Wan et al., 2024b; Jiang et al., 2023a). Furthermore, advances in culture-specific LLMs and synthetic data generation offer the potential to curate more culturally representative training and evaluation datasets beyond Western-centric narratives, supporting the development of culturally robust LLMs.

Building Conflict-Aware AI Systems As outlined in this survey, various types of conflicts can arise in a model's input, each requiring different handling depending on the task. We argue that downstream applications should not treat all conflicts uniformly; rather, responses should be tailored to the conflict type. For instance, conflicts due to ambiguity should elicit clarification questions, factual contradictions should trigger reasoning over evidence, and opinion-based disagreements should induce balanced, multi-perspective responses. Realizing such capabilities requires models to be aware of the potential conflicts and classify them according to a systematic taxonomy. Yet, current research lacks frameworks to distinguish and operationalize these conflict types. Our proposed taxonomy offers a foundational step toward enabling conflict-aware systems that can recognize, interpret, and appropriately address diverse conflicts in downstream applications.

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, challenges, and future directions.

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 Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2024. Context-dpo: Aligning language models for context-faithfulness.

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, William Yang Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

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.

Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2025. First-person fairness in chatbots. In *The Thirteenth International Conference on Learning Representations*.

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 and 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 Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495, Mexico City, Mexico. Association for Computational Linguistics.

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. 2024. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. In *The Thirteenth International Conference on Learning Representations*.

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.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023a. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023b. 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.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. ECON: On the detection and resolution of evidence conflicts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7816–7844, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2023. Analyzing dataset annotation quality management in the wild.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

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. 2021. 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, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7827–7835, Albuquerque, New Mexico. 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. 2022a. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022b. 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

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. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28657–28670.

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.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2025. Vikhr: The family of open-source instruction-tuned large language models for russian.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151, Miami, Florida, USA. Association for Computational Linguistics.

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 Tongshuang Wu. 2024. ”merge conflicts!”' exploring the impacts of external knowledge distractors to parametric knowledge graphs. In *First Conference on Language Modeling*.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 66127–66137. Curran Associates, Inc.

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.

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.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

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.

Yujie Sun, Dongfang Sheng, Zihan Zhou, and Yifei Wu. 2024. Ai hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11(1):1–14.

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.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 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. 2024a. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024b. Knowledge fusion of large language models. In *The Twelfth International Conference on Learning Representations*.

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.

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. In *First Conference on Language Modeling*.

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. In *The Twelfth International Conference on Learning Representations*.

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 Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. 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 7051–7063, Mexico City, Mexico. Association for Computational Linguistics.

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

A Summary Tables

In the summary tables, *dataset* covers prior work that proposed datasets and benchmark, *method* covers work that focus on mitigation strategies, and *analysis* presents work that aim at providing insights through experiments.

Table 1: Datasets, methods, and analysis for conflicts in natural texts

Conflict Type	**Sub-type**	**Category**	**Work**
Factual	Ambiguity	Dataset	SituatedQA (Zhang and Choi, 2021)
			AmbigQA (Min et al., 2020)
			Time-sensitive QA (Chen et al., 2021)
			StreamingQA (Liska et al., 2022)
			Real-time QA (Kasai et al., 2023)
		Method	Disambiguate then answer (Cole et al., 2023)
			Time-aware language model (Dhingra et al., 2022)
	Contradictory Evidence	Dataset	QACC (Liu et al., 2024)
			Contra-Doc (Li et al., 2024b)
			WhoQA (Pham et al., 2024)
			Machine-generated (Pan et al., 2023)
			Machine-generated (Wan et al., 2024a)
			Machine-generated (Hong et al., 2024)
			Machine-generated (Jiayang et al., 2024)
			Entity-substitution (Chen et al., 2022a)
			Rule-based (Kazemi et al., 2023)
		Method	Finetuned Calibrator (Chen et al., 2022a)
			Finetuned w/ Explanation (Liu et al., 2024)
			Finetuned discriminator (Hong et al., 2024)
		Analysis	Confirmation bias (Xie et al., 2024)
Opinion	Perspectives	Dataset	PERSPECTRUM (Chen et al., 2019)
			Multi-OpEd (Liu et al., 2021)
			NeuS (Lee et al., 2022)
			ConflictingQA (Wan et al., 2024a)
			Reddit (Plepi et al., 2024)
		Method	Multi-task learning (Liu et al., 2021)
			Opinion summarization (Jiang et al., 2023b)
			Tailored generation (Plepi et al., 2024)
	Framing Bias	Dataset	MFC (Card et al., 2015)
			GVFC (Liu et al., 2019)
			BASIL (Fan et al., 2019)
		Method	Multifaceted analysis (Milbauer et al., 2021)
			Pre-training (Liu et al., 2022b)
			Disentanglement (Chen et al., 2023)
			Multi-task learning (Lee et al., 2022)
		Analysis	Sentence-level (Lei et al., 2022)

Table 2: Datasets, methods, and analysis for conflicts in human-annotated texts

Conflict Type	Sub-type	Category	Work
Human-Annotated	Disagreement	Dataset	Twitter (Sandri et al., 2023) RLHF (Dsouza and Kovatchev, 2025) DiscoGeM (Yung and Demberg, 2025) NLI (Pavlick and Kwiatkowski, 2019)
		Method	Probabilistic model (Sheng et al., 2008) Multi-task (Mostafazadeh Davani et al., 2022) Soft labels (Uma et al., 2021) Soft labels (Fornaciari et al., 2021)
		Analysis	Survey (Uma et al., 2021) Survey (Klie et al., 2023) Offensive language (Sandri et al., 2023) NLI (Jiang and de Marneffe, 2022) Task design (Dsouza and Kovatchev, 2025) Free choice (Yung and Demberg, 2025) Personal belief (Sap et al., 2022) Demographic data (Wan et al., 2023)
	Biases	Dataset	Gender (Buolamwini and Gebru, 2018) Argument mining (Thorn Jakobsen et al., 2022) POPQUORN (Pei and Jurgens, 2023)
		Analysis	Western-centric (Faisal et al., 2022) Toxicity (Wan et al., 2023) Racist outcome (Sap et al., 2019)

Table 3: Datasets, methods, and analysis for conflicts during model interactions

Conflict Type	Sub-type	Category	Work
Knowledge Conflict	Context v.s. Memory	Dataset	Entity substitution (Longpre et al., 2021) Entity substitution (Chen et al., 2022a) Instruction-based (Wang et al., 2024) Misinformation injection (Pan et al., 2023) KRE (Ying et al., 2024) context-conflicting (Tan et al., 2024)
		Method	Data Augmentation (Longpre et al., 2021) Abstention (Chen et al., 2022a) Instruction-based (Wang et al., 2024)
	Within & Across	Analysis	LM-vs-LM fact-checking (Cohen et al., 2023) Cross-modality (Zhu et al., 2024) Intra-model contradiction (Zhao et al., 2024)
Hallucination	Factual	Dataset	TruthfulQA (Lin et al., 2022) FRANK (Pagnoni et al., 2021) q^2 (Honovich et al., 2021) Pinocchio (Hu et al., 2024) MiniCheck (Tang et al., 2024)
		Method	RAG (Lewis et al., 2020) RAG (Shuster et al., 2021) Knowledge graph (Guan et al., 2024) Disentanglement (Neeman et al., 2023) QA validation (Kryscinski et al., 2020) QA validation (Wang et al., 2020) Entailment-based (Laban et al., 2022) SelfCheckGPT (Manakul et al., 2023)
		Analysis	Less popular entities (Mallen et al., 2023)
	Contextual	Dataset	HalOmi (Dale et al., 2023) HADES (Liu et al., 2022a) RAGTruth (Niu et al., 2024)
		Method	Context-aware decoding (Shi et al., 2024) Long context (Liu et al., 2025) Context-DPO (Bi et al., 2024) CR-DPO (Huang et al., 2025)
		Analysis	Summarization (Maynez et al., 2020) Translation (Raunak et al., 2021)