Synthetic Dialogue Data Generation: A Comprehensive Survey of Methods, Evaluation, and Challenges

Anshul Chavda and **Pushpak Bhattacharyya**
Indian Institute of Technology Bombay, India
{anshulchavda,pb}@cse.iitb.ac.in

Abstract

Synthetic dialogue data generation has emerged as a pivotal area in natural language processing, driven by the escalating capabilities of large pretrained and instruction-tuned language models to simulate realistic multi-turn conversations. This survey traces the field's evolution from early template-based and rule-based systems through hierarchical encoder–decoder and transformer-based neural architectures to contemporary LLM-driven pipelines that leverage prompt engineering, agent-based-simulation, and external knowledge grounding. We examine core generation paradigms—prompt-based synthesis, agent-based simulation, and plan-and-realize grounding—analyzing how each shapes diversity, coherence, and controllability in synthetic dialogues. A comprehensive review of evaluation metrics follows, encompassing surface-level metrics (BLEU, ROUGE), embedding-based measures (BERTScore, MAUVE), dialogue-specific criteria (USR, DialogRPT, Distinct-n), and human-in-the-loop assessments. Finally, we identify key challenges—including hallucination, persona drift, lack of standardized benchmarks, and ethical/privacy concerns—and outline future research directions toward modular, interoperable architectures, unified evaluation suites, and human-centered feedback loops to enhance the scalability, reliability, and trustworthiness of synthetic dialogue generation.

1 Introduction

Dialogue systems have become an integral component of modern natural language processing (NLP), powering applications ranging from customer support and task-oriented assistants to open-domain chatbots and social dialogue agents. However, the success of data-driven dialogue models has historically depended on large-scale annotated conversational corpora, which are expensive to collect, domain-specific, and often privacy-sensitive. To address these limitations, researchers have explored methods for generating synthetic dialogue data to augment or replace human-authored examples. Early approaches leveraged handcrafted templates and rule-based user simulators (Schatzmann et al., 2007), while later neural methods employed Seq2Seq architectures (Vinyals and Le, 2015), hierarchical models (Serban et al., 2016a), and reinforcement learning for dialogue response optimization (Li et al., 2016b). Despite progress, these paradigms face challenges in scalability, diversity, and linguistic richness.

The advent of large-scale pretrained language models (LLMs), such as GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022), has marked a paradigm shift in synthetic data generation. LLMs, trained on massive text corpora, exhibit strong few-shot and zero-shot capabilities, enabling the generation of coherent, contextually appropriate dialogues via prompt-based techniques. Moreover, recent work on instruction tuning and chain-of-thought prompting has improved the controllability and reasoning abilities of LLMs (Wei et al., 2022). Leveraging LLMs to synthesize task-oriented dialogs from seed instructions (Wang et al., 2023) or prompt-based persona conditioning (Chen et al., 2023) has produced high-quality synthetic dialogues that rival human-authored data. These developments open new avenues for scalable, domain-agnostic dialogue synthesis while raising novel challenges in data quality, consistency, and evaluation.

In this survey, we provide a comprehensive overview of LLM-driven synthetic dialogue dataset generation. Our contributions are:

- We examine the core synthetic dialog generation methodologies—prompt-based synthesis, LLMs as agents, and plan-and-realize grounding—highlighting how each approach shapes dataset quality and diversity. (Section 4)

- We survey evaluation metrics used for synthetic dialogue corpora, covering automatic surface and embedding-based metrics (BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020a), MAUVE (Pillutla et al., 2021)), dialogue-specific measures (USR (Mehri and Eskenazi, 2020), DialogRPT (Gao et al., 2020), Distinct-n (Li et al., 2016a)), and human judgment practices. (Section 5)

- Identify and discuss critical challenges—factual consistency, controllability, scalability, and ethical/privacy issues—and propose promising future research directions toward modular architectures and standardized benchmarks. (Section 6)

The remainder of this paper is organized as follows. Section 2 introduces core definitions and dialogue types. Section 3 briefly reviews pre-LLM dialogue generation paradigms. Section 4 dissects generation pipelines in detail, discussing their motivations, methodologies, and artifacts. Section 5 surveys evaluation metrics used for synthetic dialogue quality. Section 6 discusses key challenges with LLM-based synthetic dialog data generation. Finally, Section 7 offers summary reflections, concluding remarks and future directions.

2 Background

2.1 What Is "Synthetic Dialogue Data"?

We define *synthetic dialogue data* as any conversational corpus whose turns, whether they are user utterances, system responses, or both, are generated automatically rather than captured directly from human to human or human to machine interactions. Synthetic data aims to emulate the functional properties of real conversations (e.g., turn-taking, grounding, error patterns) while offering advantages in scale, controllability, and privacy.

Unlike classic data augmentation which is focused on producing variants of existing utterances, synthetic dialogue data can instantiate entirely novel conversation flows, including new slot combinations, dialogue acts, and error patterns. Compared to purely human-collected corpora, synthetic datasets offer:

- **Unlimited Scale:** Generation can be repeated to meet any desired dataset size.

- **Domain Control:** Parameters (e.g., slot distributions, error rates) can be explicitly tuned to target specific applications or stress-test edge cases.

- **Privacy Preservation:** By avoiding direct use of sensitive user logs, synthetic data reduces risks associated with personally identifiable information and compliance with data-protection regulations.

- **Rapid Domain Prototyping:** New domains or languages can be bootstrapped with minimal human annotation via cross-lingual back-translation or multilingual prompts.

2.2 Dialogue Types and Task Settings

Synthetic dialogue data can be customized for different types of conversations and use cases. We summarize the key settings where such data is commonly applied:

Task-Oriented Dialogue: These are goal-driven conversations, such as booking a flight, reserving a restaurant, or checking the weather. Each turn aims to fill specific *slots* (e.g., `restaurant_name`, `time`, `location`) that help accomplish the user's task. Simulators for task-oriented dialogue often generate synthetic conversations by exhaustively combining user intents and slot values. Reinforcement learning can also be used to train system policies that maximize success metrics like task completion or user satisfaction. Figure 1 shows an example of a restaurant booking scenario where the system identifies the user goal and responds accordingly. This type of interaction is typical in multi-turn systems trained via templates, simulators, or large language models.

Open-Domain (Chit-Chat): These are open-ended, social interactions with no specific goal other than engagement and naturalness. For instance, a chatbot discussing weekend plans or giving casual movie recommendations. Synthetic generation in this setting typically uses large language models (LLMs) such as GPT-3 or BlenderBot, which are prompted with examples or personas to maintain consistency and variety (Zhang et al., 2020b; Roller et al., 2021; Brown et al., 2020). Figure 2 shows an example

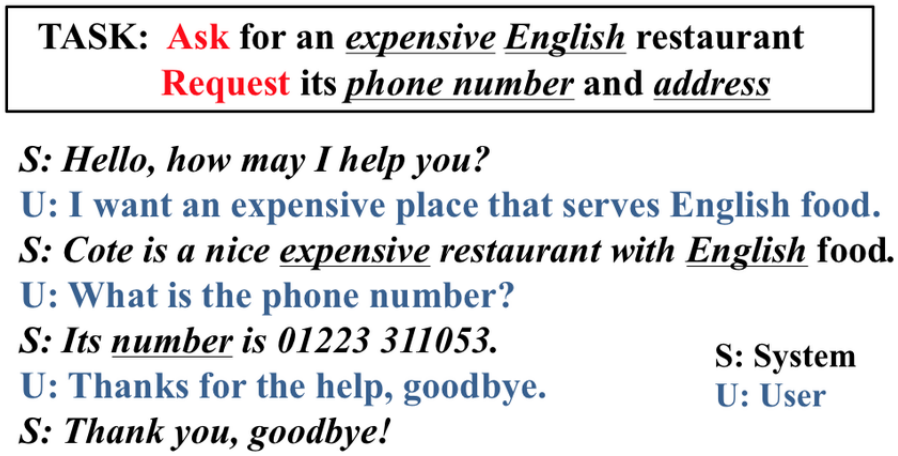

Figure 1: Example of a task-oriented dialogue where the user makes specific requests (e.g., find an expensive English restaurant and retrieve its phone number).

of a persona-driven open-domain dialogue, where the agent maintains consistency with its background profile throughout the conversation.

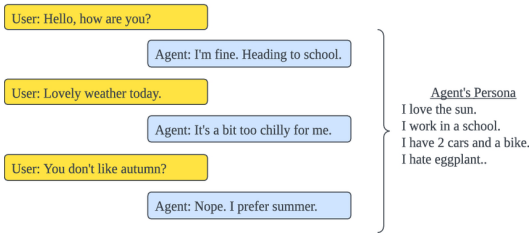

Figure 2: An example of a persona-conditioned open-domain chatbot where the agent's responses are influenced by a predefined persona.

Mixed-Initiative Dialogue: In mixed-initiative dialogues, both the user and the system take turns leading the interaction. For example, if the user says "I want to eat out," the system might ask "Do you prefer indoor or outdoor seating?" before proceeding. As shown in Figure 3, the system actively probes for more context while allowing the user to express emotional and narrative content, requiring careful management of dialogue flow and sensitivity.

Multimodal Dialogue: Some conversations involve more than just text—they include vision (images, scenes), speech (prosody, pauses), or gestures. For instance, a dialogue system helping someone navigate a physical environment or describing a photo. Synthetic data for multimodal dialogue pairs generated utterances

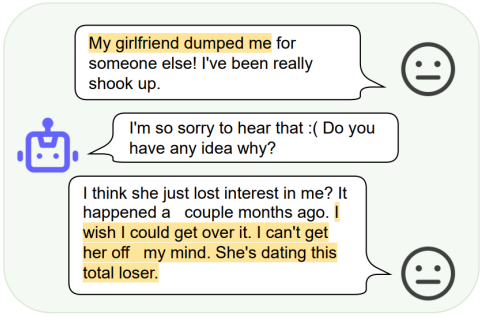

Figure 3: Example of a mixed-initiative dialogue where the system responds empathetically and prompts further disclosure, illustrating fluid turn-taking.

with contextual signals like images or speech features, enabling the training of embodied or speech-aware conversational agents. Figure 4 illustrates a multimodal shopping assistant that integrates both textual conversation and image-based product displays, enabling fine-grained, context-aware interaction.

3 Overview of Synthetic Dialogue Generation Paradigms

This section provides a concise historical context for synthetic dialogue generation, highlighting major paradigms prior to the dominance of large language models.

- **Template-based and Rule-Based Generation.** Early conversational agents relied on handcrafted templates and production rules to generate responses. Systems such as ELIZA (Weizenbaum, 1966) and A.L.I.C.E. (Wallace, 2009) demonstrated the

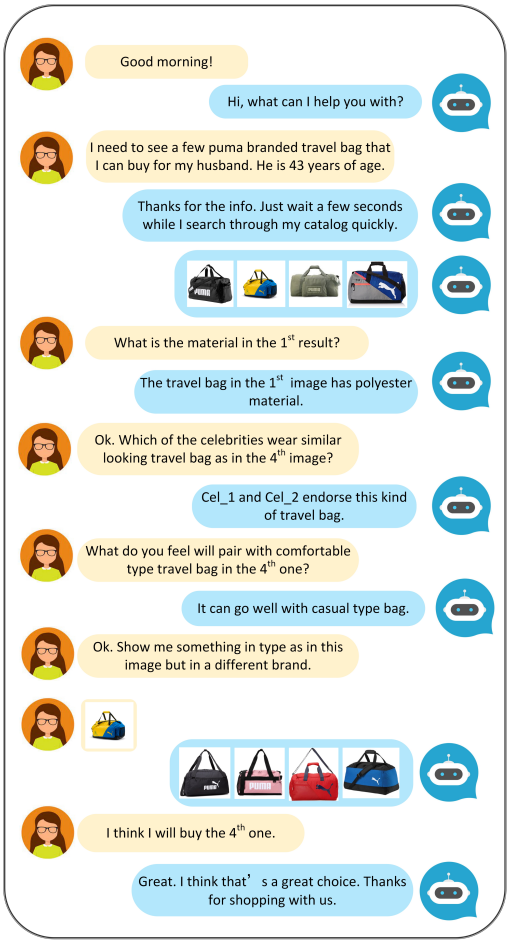

Figure 4: An example of a multimodal dialogue system where the assistant responds to user queries using both text and visual product options.

feasibility of dialogic interaction through pattern matching and scripted reply templates. As shown in Figure 5, these systems generated responses by transforming user inputs according to predefined rules. Despite their interpretability and control, these methods suffer from limited linguistic diversity and brittle behavior outside predefined scenarios.

- **Simulation-Based Methods.** To support policy learning in task-oriented systems, researchers developed user simulators that mimic human interlocutors. Agenda-based simulators employ a goal stack and slot-filling agenda to generate user utterances, enabling reinforcement learning of dialogue policies (Schatzmann et al., 2007). Probabilistic simulators further introduced stochastic variation but remained confined to limited domains and ontologies (Wen et al., 2015).

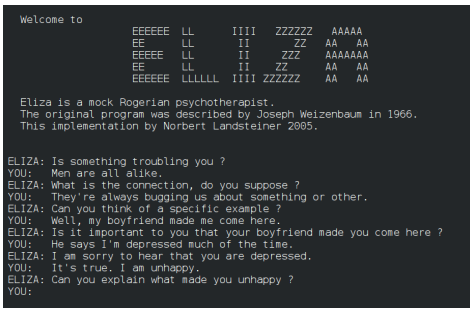

Figure 5: A sample conversation with ELIZA, an early rule-based chatbot. The dialogue demonstrates ELIZA's pattern-matching responses, such as rephrasing user inputs (e.g., *"He says I'm depressed"* → *"I am sorry to hear that you are depressed"*) and prompting for elaboration (e.g., *"Can you think of a specific example?"*).

- **Pre-LLM Neural Generative Models.** The advent of neural architectures ushered in data-driven generation. Hierarchical encoder-decoder models capture multi-turn context via latent variables and recurrent structures (Serban et al., 2016b). As illustrated in Figure 6, these models process dialogue turns hierarchically, with utterance-level encoding feeding into conversation-level context modeling. Subsequently, transformer-based fine-tuned models like DialoGPT leverage large-scale pretraining and self-attention to produce more coherent replies (Zhang et al., 2020b). While these models improved fluency and context tracking, their generation quality hinged on the size and domain coverage of supervised corpora.

Conventional synthetic generation approaches face scalability challenges: template-based methods lack diversity; simulators require painstaking design; and neural models demand extensive annotated data for finetuning. Moreover, these paradigms often struggle with open-domain dialogues, long-range coherence, and dynamic adaptation to new topics, motivating the shift toward promptable LLMs.

4 LLM-Driven Synthetic Dialogue Data Generation

In this section, we present a detailed review of prominent methods that leverage large language models (LLMs) to construct synthetic dialogue datasets. Each method is discussed through its motivation, generation methodology, and dataset characteristics. The goal is to understand how dif-

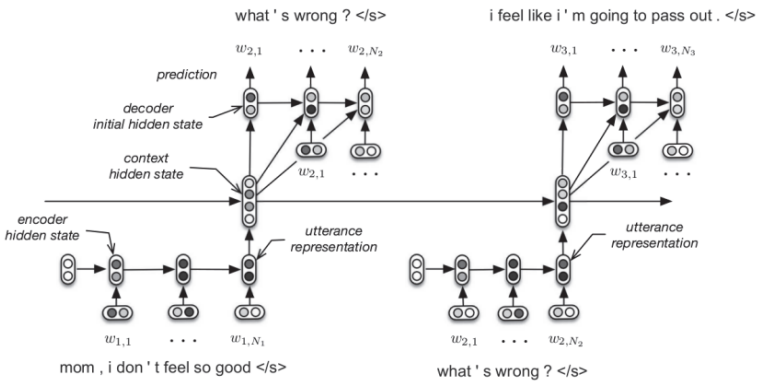

Figure 6: The computational graph of the HRED architecture for a dialogue composed of three turns. Each utterance is encoded into a dense vector and then mapped into the dialogue context, which is used to decode (generate) the tokens in the next utterance. The encoder RNN encodes the tokens appearing within the utterance, and the context RNN encodes the temporal structure of the utterances appearing so far in the dialogue. Adapted from (Serban et al., 2016b).

ferent prompting, planning, or agent simulation strategies are used to construct multi-turn conversational corpora across diverse domains.

4.1 PLACES: Prompting Language Models for Social Conversation Synthesis (Chen et al., 2023)

Motivation The PLACES framework addresses the challenges of collecting high-quality conversational data for social dialogue systems. Traditional approaches relying on crowdworkers are expensive and yield inconsistent quality, while existing datasets often lack diversity or are limited to dyadic interactions. PLACES proposes using expert-written examples to guide LLMs in generating synthetic conversations that match the quality of human-collected datasets while enabling control over conversation topics and participant backgrounds.

Methodology The approach involves three key components:

- **Expert-written examples**: A small pool of 10 high-quality conversations written by experts, each accompanied by a "recipe" specifying the topic and background information for participants (e.g., "Alice loves cats. Bob is more of a dog person.").
- **Few-shot prompting**: For generation, the model (OPT-30B (Zhang et al., 2022)) is prompted with 3 randomly sampled example conversations plus a new target recipe.
- **Multi-party extension**: The method is extended to triadic conversations by adding a third participant's background information.

Key technical details:

- Generates 5,592 conversations matching the topic-subtopic pairings from the FITS dataset (Xu et al., 2023)
- Average conversation length: 9.29 turns (12.84 words/turn)

As shown in Figure 7, the model generates a contextually appropriate dialogue based on the given topic and background.

Evaluation and Results The synthetic conversations were rigorously evaluated through:

- **Human assessment** (28 crowdworkers rating 200 samples):
  - Rated more coherent (4.61/5) and consistent (4.63/5) than DailyDialog (Li et al., 2017)
  - 95% matched their prescribed topics
- **Lexical diversity**: Higher Distinct-N scores than human datasets for bigrams/trigrams
- **Fine-tuning results**: BlenderBot models (Roller et al., 2021) trained on synthetic data performed comparably to those trained on human data in interactive evaluations
- **Multi-party conversations**: Synthetic triadic dialogues outperformed human-collected datasets across all quality dimensions

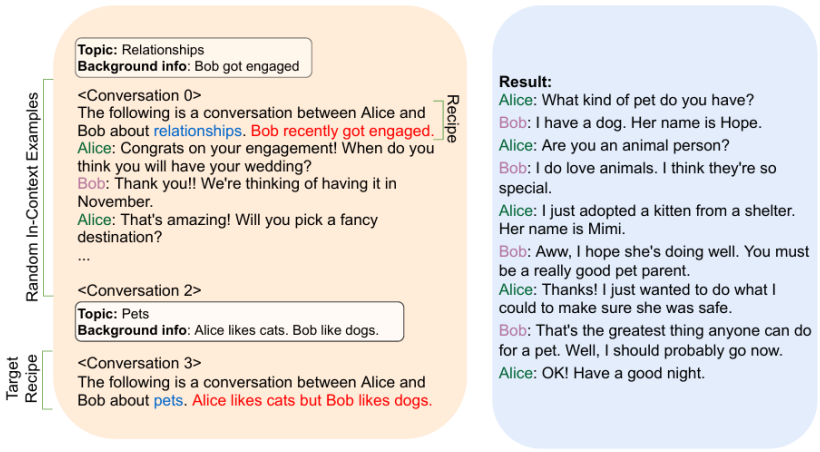

Figure 7: PLACES generation framework showing the prompt structure (left) with example recipes and generated conversation (right). The model uses few-shot examples to produce topic-grounded dialogues while maintaining speaker consistency.

4.2 Synthetic Dialogue Generation Using LLM Agents (Abdullin et al., 2023)

Motivation This work addresses the specific challenge of generating training data for goal-oriented conversational agents in mathematical optimization domains. The authors focus on creating synthetic dialogues that can help users formulate linear programming (LP) problems through natural language interactions, addressing a key bottleneck where non-experts struggle to translate real-world problems into formal optimization models.

Methodology The framework employs a dual-agent architecture using GPT-4:

- A *Question Generation (QG) Agent* acts as the conversational assistant, systematically eliciting LP problem components (decision variables, objective function, constraints) through targeted questions
- A *Question Answering (QA) Agent* simulates the user, responding based on problem descriptions from the NL4Opt dataset (Ramamonjison et al., 2022)
- The dialogue terminates when the QG agent produces a complete summary of the LP problem, verified against the original description

Key innovations include:

- Domain-specific prompt engineering to maintain mathematical accuracy while avoiding technical jargon
- An automatic summary verification mechanism to ensure dialogue quality
- Temperature-controlled generation to balance consistency and diversity

Dataset and Evaluation The authors generated 476 dialogues (with 28 human-annotated examples) featuring:

- Average length of 20 turns (3,658 characters)
- 97% success rate in producing valid LP problem summaries
- Coverage of 9 common constraint types from real-world optimization problems

Evaluation combined:

- *Human assessment* (4 annotators) showing high scores for information recall (4.29/5) and precision (4.38/5)
- *Automatic metrics* with strong BERTScore (Zhang et al., 2020a) and moderate ROUGE-L (Lin, 2004) performance
- A novel GPT-4 evaluator achieving fair correlation of 0.67 with human precision judgments

4.3 SynDG: Grounded Dialogue via Plan-and-Realize Framework (Bao et al., 2023)

Motivation Existing approaches for generating knowledge-grounded dialogues often lack explicit modeling of dialogue flow—the structured progression of topics that ensures conversation coherence.

Human dialogues naturally transition between related knowledge pieces (e.g., moving from "husky dogs" to "sled dogs" to "huskies as pets"). Current synthetic generation methods either require expensive human annotation or produce incoherent outputs by neglecting this flow structure. SynDG addresses this by automating high-quality dialogue generation while preserving logical knowledge progression.

Methodology The framework (as shown in Figure 8) employs a three-stage pipeline:

- **Flow Construction**: Heuristic sampling from knowledge sources (Wikipedia articles/persona profiles) creates topic sequences respecting domain-specific patterns. For WoW (Wizard of Wikipedia) (Dinan et al., 2019), 90% of wizard turns sample from central topic sentences.
- **Utterance Realization**: A T5-Large (Raffel et al., 2020) model incrementally generates each utterance conditioned on:
  - Previous dialogue history
  - Current knowledge piece (marked with [t] tags)
  - 1–2 subsequent knowledge pieces (for coherence)
- **Quality Filtering**: Two T5-based scorers evaluate:
  - Flow-level consistency (masked knowledge prediction)
  - Utterance-level coherence (masked utterance prediction)

  Bottom 50% scored dialogues are discarded.

Dataset The framework produces:

- **36,860 WoW (Wizard of Wikipedia)-style (Dinan et al., 2019) dialogues** (18,430 retained after filtering)
- **6,600 PersonaChat (Zhang et al., 2018) dialogues** (from 10k initial samples)

Key statistics:

- Avg. 10 turns/dialogue (WoW (Dinan et al., 2019)) and 16 turns (PersonaChat (Zhang et al., 2018))
- 52,800 training samples for PersonaChat
- Covers both seen/unseen topics (WoW (Dinan et al., 2019) Test Unseen shows 14.67 BLEU-4)

Data quality is validated by:

- Human evaluation (30%+ preference over baselines)
- Low-resource experiments (1/16 data matches full-data performance)

The reviewed approaches showcase a diverse toolkit for synthetic dialogue generation, including prompt-based generation (Chen et al., 2023), agent-based simulation (Abdullin et al., 2023), and grounded realization (Bao et al., 2023). Each method contributes unique design trade-offs and artifacts, enriching the dialogue modeling landscape.

5 Evaluation Metrics for Synthetic Dialogue Data

Evaluating synthetic dialogue datasets require careful consideration of metrics that reflect not only the surface-level similarity of generated text to references but also dialogue-specific qualities such as contextual coherence, diversity, engagement, and faithfulness to external knowledge. In this section, we review automatic metrics, dialogue-specific metrics, and human evaluation tailored to synthetic dialogue data.

5.1 Automatic Metrics

Automatic metrics provide rapid, reproducible assessments of synthetic dialogues, facilitating large-scale comparisons.

5.1.1 N-gram Overlap Metrics

BLEU (Papineni et al., 2002) computes the precision of n-gram matches between generated utterances and reference dialogues. Despite its widespread use, BLEU often fails to capture semantic adequacy in dialogue due to high lexical variability.

ROUGE (Lin, 2004) emphasizes recall of overlapping n-grams or sequences, making it suitable for capturing information coverage but similarly limited by surface matching.

METEOR (Banerjee and Lavie, 2005) improves on BLEU (Papineni et al., 2002) by incorporating synonym matching, stemming, and a penalty for fragmentation, yielding better correlation with human judgments in short text generation tasks.

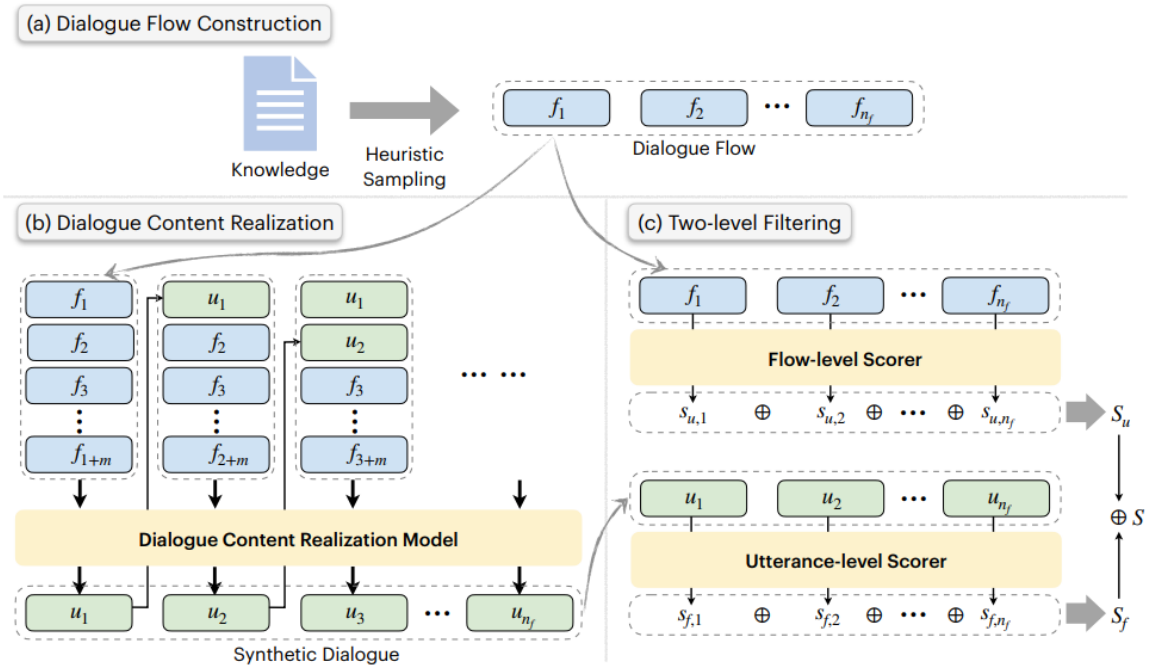

Figure 8: SynDG's generation pipeline: (a) Knowledge-guided flow construction, (b) Incremental utterance realization with future knowledge window, (c) Two-stage quality filtering.

5.1.2 Embedding-based Metrics

Moving beyond exact matches, embedding-based metrics assess semantic similarity.

BERTScore (Zhang et al., 2020a) computes token-level cosine similarities using contextual embeddings from pretrained transformers, showing higher correlation with human ratings for dialogue fluency and relevance.

MAUVE (Pillutla et al., 2021) measures the divergence between distributions of real and generated text in embedding space, quantifying how closely synthetic dialogues match human conversational patterns.

5.1.3 Diversity Metrics

Synthetic dialogue data must avoid generic or repetitive utterances.

Distinct-n (Li et al., 2016a) calculates the ratio of unique n-grams (typically unigrams and bigrams) to total generated tokens, with higher values indicating greater lexical diversity.

5.1.4 Task-specific F1

For synthetic data aimed at question–answer or slot-filling dialogues, **F1** measures the harmonic mean of precision and recall of key information (e.g., slot values) extracted from generated utterances, aligning evaluation with downstream task objectives.

5.2 Dialogue-Specific Metrics

Automatic metrics borrowed from general NLG often overlook salient dialogue properties. Specialized metrics have been developed to address these gaps.

5.2.1 Context Coherence

Metrics like **USR** (UnSupervised and Reference-free evaluation) (Mehri and Eskenazi, 2020) predict coherence scores by modeling the likelihood of a response given dialogue history, without requiring references.

5.2.2 Engagement

Engagement Predictors estimate how likely a dialogue turn is to stimulate user interest. For instance, the **EngageScore** (Ghazvininejad et al., 2017) uses reinforcement learning feedback signals to approximate user engagement, while learned metrics such as **DialogRPT** (Gao et al., 2020) rank responses by preference models trained on upvotes and downvotes from large-scale human interactions.

5.2.3 Topical Diversity

Evaluating whether synthetic dialogues cover a breadth of topics can be quantified by computing the entropy over topic assignments, using pretrained topic classifiers (Dieng et al., 2020). High

entropy indicates a wider topical spread in the dataset.

5.3 Human Evaluation

Human judgments remain the gold standard for assessing dialogue quality. Crowdsourced annotators or expert raters evaluate synthetic dialogues along multiple axes.

5.3.1 Fluency

Annotators rate the grammaticality and readability of each utterance, typically on a Likert scale (1–5). Fluency scores correlate with automatic perplexity measures but capture nuanced linguistic errors beyond n-gram patterns.

5.3.2 Relevance and Coherence

Human judges assess whether each response appropriately follows the preceding context. Liu et al. (2022) recommend pairing Likert-scale relevance judgments with binary coherence flags to disentangle topicality from turn-level consistency.

5.3.3 Informativeness

For grounded or task-oriented synthetic dialogues, evaluators judge whether utterances convey correct and useful information. This can involve checking factual accuracy against a knowledge source or verifying slot completion in a task schema.

5.4 Faithfulness in Retrieval-Augmented Generation

For synthetic dialogues that incorporate external knowledge via retrieval (RAG), specialized metrics assess faithfulness and attribution.

5.4.1 Attribution Accuracy

Evaluating whether generated content correctly cites or references retrieved documents, measured by the proportion of facts in outputs traceable to source passages (Lewis et al., 2020).

5.4.2 Groundedness Score

Metrics such as **Groundedness** (Dziri et al., 2022) compute the semantic alignment between generated utterances and retrieved knowledge, using embedding-based similarity thresholds to ensure the model remains anchored in the retrieval.

6 Challenges

As synthetic dialogue generation methods mature, a number of critical challenges have emerged that must be addressed to fully realize the potential of large-scale, high-quality conversational corpora. In this section, we identify key obstacles spanning data quality, model controllability, evaluation, ethics, and deployment.

6.1 Data Quality and Fidelity

6.1.1 Hallucination and Factual Inconsistency

Despite remarkable fluency, LLM-based dialogue generators often produce *hallucinated* content—responses that are grammatical but factually incorrect or unverifiable. This is especially problematic for grounded and task-oriented settings where user trust hinges on accuracy (e.g., booking systems, medical advice bots). Current mitigation strategies include retrieval-augmented generation (RAG) and explicit grounding prompts, yet these approaches struggle with:

- **Source attribution:** Systems may reference knowledge that is not present in the retrieval index, or fail to properly cite the origin of facts.
- **Controlled integration:** Seamlessly interleaving retrieved facts with generated context without breaking discourse coherence remains an open problem.
- **Hallucination detection:** Automated metrics for flagging hallucinations in conversational settings are underdeveloped, making large-scale filtering unreliable.

6.1.2 Diversity versus Quality Trade-off

High lexical and topical diversity often comes at the expense of coherence or task success. Systems with aggressive sampling (high temperature, nucleus sampling with large p) produce varied utterances but risk incoherent tangents; conservative decoding yields safe yet repetitive dialogues.

- **Balancing exploration and exploitation:** Current temperature and nucleus thresholds are hand-tuned and do not generalize across domains.
- **Adaptive decoding:** Few approaches dynamically adjust decoding hyperparameters mid-dialogue based on context or user feedback.

6.1.3 Scalability of Generation Pipelines

LLM-driven pipelines require substantial compute, particularly when synthesizing millions of multi-turn dialogues. Efficiency bottlenecks arise in:

- **Prompt efficiency:** Large few-shot prompts consume context window and degrade throughput.
- **Agent-based simulation:** Multi-agent enviroments multiply inference cost quadratically with number of participants.
- **Filtering and verification:** Two-stage quality filters (flow-level, utterance-level) require additional forward passes, doubling or tripling inference time.

6.2 Controllability and Personalization

6.2.1 Persona and Style Drift

Maintaining consistent speaker persona or dialogue style over lengthy interactions remains challenging. Drift occurs when:

- **Prompt leakage:** Few-shot context no longer fully represents persona after many turns.
- **Implicit biases:** LLMs introduce sociolinguistic biases that override explicit persona attributes.

6.2.2 Fine-Grained Control Over Dialogue Acts

Synthetic dialogues often lack explicit annotation of dialogue acts (e.g., question, request, affirmation), making it difficult to train controllable dialogue managers. While some plan-and-realize frameworks build topic flows, they rarely incorporate discourse-level act planning.

6.3 Evaluation and Benchmarking

6.3.1 Lack of Standardized Benchmarks

Synthetic dialogue research employs diverse datasets and custom metrics, impeding fair comparison across methods. Existing benchmarks often focus on surface metrics or small-scale human studies.

6.3.2 Human Evaluation Scalability

Crowdsourced evaluations are costly and lack inter-annotator consistency. Moreover, evaluations rarely reflect real user interactions.

6.4 Ethical, Privacy, and Bias Considerations

6.4.1 Propagation of Societal Biases

LLMs trained on web corpora inherit biases (gender, racial, ideological) that can manifest in synthetic dialogues, potentially amplifying harmful stereotypes.

6.4.2 Privacy and Data Leakage

Synthetic data is often promoted as privacy-preserving, yet LLMs can memorize and regurgitate training data. The risk of exposing sensitive content—especially when fine-tuned on private logs—remains real.

6.5 Domain Adaptation and Low-Resource Scenarios

6.5.1 Multilingual and Cross-Lingual Generation

Most synthetic dialogue work focuses on English, leaving non-English languages underrepresented. Prompt engineering and in-context learning often fail in low-resource languages.

6.5.2 Specialized Domains and Jargon

Technical domains (law, medicine, finance) require precise terminology and adherence to domain conventions. Off-the-shelf LLMs lack deep domain expertise, leading to superficial or misleading dialogues.

6.6 Interactive and Continual Learning

Real-world dialogue agents must adapt over time based on user feedback, evolving language trends, and shifting user needs. Static synthetic corpora cannot capture this dynamism.

6.7 Robustness, Safety, and Deployment

6.7.1 Adversarial and Safety-Critical Scenarios

Dialogue systems in healthcare, finance, or autonomous vehicles must be robust to adversarial inputs and safe-fail under uncertainty.

6.7.2 Real-Time Constraints and Edge Deployment

Large models typically run in the cloud, incurring latency and privacy concerns. Edge deployment demands lightweight yet capable systems.

6.8 Synthesis of Directions

The challenges outlined above are deeply interconnected. For example, improving factuality via retrieval can also aid diversity by grounding prompts in heterogeneous knowledge sources; enhancing controllability of dialogue acts supports both evaluation rigor and user trust; and deploying compact agents on edge devices demands advances in model compression, interactive learning, and safety. We

therefore advocate for research that bridges these dimensions through:

- **Modular, interoperable pipelines:** Architectures where retrieval, planning, generation, filtering, and evaluation are decoupled yet communicate via shared interfaces.
- **Unified benchmarks:** Large-scale, multilingual, multi-domain evaluation suites that reflect real-world usage patterns and safety constraints.
- **Human-centered feedback loops:** Systems seamlessly integrate diverse forms of human input—preferences, corrections, and safety judgments—to close the loop on synthetic data quality.

By tackling these challenges in an integrated manner, the field can move toward the generation of synthetic dialogue corpora that not only scale arbitrarily but also uphold the fidelity, diversity, safety, and ethics required for next-generation conversational AI.

7 Summary, Conclusion and Future Work

7.1 Summary

This survey has traced the evolution of synthetic dialogue data generation from early template-based and rule-based methods to modern pipelines driven by large language models (LLMs). We examined three major paradigms. First, prompt-based synthesis, such as PLACES (Chen et al., 2023), uses few-shot prompting and persona-topic conditioning to generate diverse and coherent social conversations. Second, agent-based simulation employs LLMs in self-play settings where one model plays both user and system roles, enabling the creation of task-oriented dialogues without requiring real user data (Abdullin et al., 2023). Third, plan-and-realize grounding approaches, such as SynDG (Bao et al., 2023), use structured topic flows or external knowledge as scaffolds before realizing utterances with a generation model and applying multi-stage filtering for quality assurance. Evaluation of synthetic dialogues remains a challenge. We reviewed common automatic metrics like BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020a), and MAUVE (Pillutla et al., 2021), as well as dialogue-specific metrics including USR (Mehri and Eskenazi, 2020), Distinct-n (Li et al., 2016a), and DialogRPT (Gao et al., 2020). Human evaluations typically assess fluency, coherence, and informativeness, while recent work also emphasizes factual consistency in retrieval-augmented generation (Lewis et al., 2020; Dziri et al., 2022). These efforts have facilitated the creation of large-scale synthetic dialogue corpora that significantly enhance conversational model performance.

7.2 Conclusion

Synthetic dialogue generation has progressed rapidly, shifting from hand-crafted, rule-based techniques to sophisticated LLM-driven pipelines capable of producing vast and diverse corpora. These modern approaches, particularly prompt-based generation (Chen et al., 2023), agent-based simulation (Abdullin et al., 2023), and plan-and-realize systems (Bao et al., 2023), have enabled the creation of high-utility datasets across open-domain and task-oriented scenarios. However, challenges remain. Ensuring factual accuracy, controlling stylistic consistency and persona fidelity, and building robust, reproducible evaluation protocols are unresolved issues. Moreover, ethical concerns such as bias propagation and privacy leakage demand careful consideration. Addressing these challenges will require collaborative, interdisciplinary efforts focused on building modular, controllable, and ethically-aware synthetic data generation frameworks. These future systems should integrate real-world feedback and support transparent benchmarking to advance trustworthy and inclusive conversational AI.

7.3 Future Work

To overcome the current limitations of synthetic dialogue generation, future research should prioritize several directions. First, hybrid retrieval–generation architectures should be developed to combine external knowledge grounding with the fluent generation capabilities of LLMs, thereby reducing hallucinations and improving factual consistency. Second, generation pipelines should decouple dialogue act planning from surface realization, enabling finer-grained control over conversational flow and function. Third, interactive and continual learning strategies must be integrated, allowing systems to adapt in real time to user feedback through techniques such as reinforcement learning with human preferences (RLHF) and active learning. Fourth, the field urgently needs universal benchmarking frameworks. These should in-

clude multilingual and multi-domain datasets, track hyperparameter settings, and combine standard automatic metrics (Papineni et al., 2002; Zhang et al., 2020a; Pillutla et al., 2021) with dialogue-specific evaluations (Mehri and Eskenazi, 2020; Li et al., 2016a) and human assessments. Fifth, ethical and privacy considerations should be embedded into generation workflows via differential privacy techniques, adversarial robustness testing, and automatic bias detection and mitigation strategies. Finally, advances in model compression and split-inference architectures are needed to deploy capable dialogue systems on edge devices where latency, privacy, and resource constraints are critical. Addressing these directions in a cohesive manner will ensure the generation of synthetic dialogue datasets that are scalable, high-quality, and aligned with societal values.

References

Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using LLM agents. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

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

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

Patrick 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Rindranirina Ramamonjison, Timothy Yu, Raymond Li, Haley Li, Giuseppe Carenini, Bissan Ghaddar, Shiqi He, Mahdi Mostajabdaveh, Amin Banitalebi-Dehkordi, Zirui Zhou, and Yong Zhang. 2022. Nl4opt competition: Formulating optimization problems based on their natural language descriptions. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 189–203. PMLR.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer Netherlands, Dordrecht.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.