

GODEL: Large-Scale Pre-Training for Goal-Directed Dialog

Baolin Peng[†] Michel Galley[†] Pengcheng He[†] Chris Brockett[†]
Lars Liden[†] Elnaz Nouri[†] Zhou Yu[‡] Bill Dolan[†] Jianfeng Gao[†]

[†] Microsoft Corp. [‡] Columbia University

{bapeng, mgalley, penhe, chrisbkt, laliden, elnouri, billdol, jfgao}@microsoft.com
zy2461@columbia.edu

Abstract

We introduce GODEL (Grounded Open Dialogue Language Model), a large pre-trained language model for dialog. In contrast with earlier models such as DialoGPT, GODEL leverages a new phase of *grounded* pre-training designed to better support adapting GODEL to a wide range of downstream dialog tasks that require information external to the current conversation (*e.g.*, a database or document) to produce good responses. Experiments against an array of benchmarks that encompass task-oriented dialog, conversational QA, and grounded open-domain dialog show that GODEL outperforms state-of-the-art pre-trained dialog models in few-shot fine-tuning setups, in terms of both human and automatic evaluation. A novel feature of our evaluation methodology is the introduction of a notion of utility that assesses the *usefulness* of responses (extrinsic evaluation) in addition to their communicative features (intrinsic evaluation). We show that extrinsic evaluation offers improved inter-annotator agreement and correlation with automated metrics. Code and data processing scripts are publicly available.¹

1 Introduction

This work describes the development of a very large pre-trained dialog model – Grounded Open Dialogue Language Model (GODEL). As the name indicates, GODEL is designed for general-domain conversation and is fully open-sourced. GODEL should be of technical interest for two reasons. First, it is pre-trained in three phases, successively folding in data from web text, publicly-available dialog (*e.g.*, Reddit), and a collection of existing corpora that support grounded dialog tasks. The grounded dialog corpora, which include MS MARCO (Nguyen et al., 2016) and DSTC7 (Yoshino et al., 2019), allow for more effective fine-tuning on dialog tasks where responses must

be conditioned on information external to the current conversation (*e.g.*, a retrieved document.) Second, GODEL is validated on a utility-driven suite of benchmarks specifically designed for few-shot fine-tuning of *open-ended goal-directed general-domain dialog* models. We will show that GODEL, as validated using this methodology, is more readily amenable to fine-tuning for goal-directed dialog tasks than other large pre-trained language models.

Our approach seeks to address a long-standing obstacle to general-purpose open-ended conversation models, namely a lack of robust automated evaluation criteria that can drive development (Gao et al., 2019). Recent state-of-the-art models that leverage large PLMs (*e.g.*, Zhang et al., 2019b; Freitas et al., 2020; Roller et al., 2021; Bao et al., 2021; Thoppilan et al., 2022; Gao et al., 2022) offer the potential for substantive open-ended conversational interactions, yet they resist meaningful comparison owing to the lack of consensus on evaluation.

This poses a fundamental question: what do we want of a good general-purpose dialog model in the first place? We take it as a given that it should be fluent and socially engaging. Indeed, most SOTA PLMs are primarily evaluated on such *intrinsic* communicative dimensions. But beyond that, we must also acknowledge that machine-human conversation typically serves a purpose and aims to fulfill one or more goals on the part of the user. In other words, the model must offer *utility* to the user. It is this *extrinsic* dimension of functional utility, we suggest, that constitutes the proper focus of automated evaluation in general-domain models.

In the second half of this paper (Sections 5 and 6), we explore this notion of Utility in fine-tuning GODEL on four established tasks that cover multiple domains and conversational settings, ranging from the task-oriented MultiWOZ (Budzianowski et al., 2018) to more open-ended goal-oriented tasks, *i.e.*, CoQA (Reddy et al., 2019), Wizard of Wikipedia (Dinan et al., 2018), and Wiz-

¹<http://aka.ms/GODEL>

ard of the Internet (Komeili et al., 2021). We show that GODEL improves on a baseline DIALOGPT (Zhang et al., 2019b), with performance gains in goal-directed tasks. Our results on these diverse tasks speak to the effectiveness of our approach. We also find that the focal shift to goal-directedness and extrinsic evaluation affords better inter-rater agreement and human-metric correlations.

As baselines for future research, we release three versions of GODEL: small (GODEL_B) and large (GODEL_L) initiated from T5 and T5-Large respectively, and a very large model (GODEL_{XL}^{GPT-J}) that is trained off GPT-J (Gao et al., 2020; Wang and Komatsuzaki, 2021).² We also furnish code to download datasets for fine-tuning and evaluation. The release is fully open sourced (models, data, and code) and tightly integrated with the latest version of HuggingFace’s transformers library.

2 Related Work

Large pre-trained language models (PLMs) for dialog (Zhang et al., 2019b; Freitas et al., 2020; Roller et al., 2021; Peng et al., 2021; Bao et al., 2021; Thoppilan et al., 2022) have had significant impact in Conversational AI, as these models have brought strong performance gains in competitions such as DSTC (Gunasekara et al., 2020) and ConvAI (Aliannejadi et al., 2020), and have been widely adopted in both task-oriented and chitchat research. While the applications of these dialog models are numerous, empirical evaluation of dialog pre-training has mostly been restricted to *intrinsic* evaluation (*i.e.*, agnostic to any notion of utility). For example, DialoGPT (Zhang et al., 2019b) offered Reddit-based pre-trained models that are evaluated on Relevance, Informativeness, and Humanness and various string overlap based automated evaluation metrics. Meena (Freitas et al., 2020)’s evaluation is focused on Sensibleness and Specificity, while also offering a new automated evaluation (SSA) that measures these two qualities. BlenderBot (Roller et al., 2021) contributes a variety of models that capitalize on skills including personality, empathy, and knowledge, but evaluation in this work centers mostly on measures of Humanness and Engagingness.

²The discussion of our largest model in this paper relates to GODEL_{XL} which is based on GPT-3, but which we are unable to release owing to licensing restrictions. We therefore release GODEL_{XL}^{GPT-J} a substitute model based on GPT-J that performs similarly. A comparison of GODEL models initialized from GPT-3 and GPT-J is provided in the Appendix.

Plato-XL (Bao et al., 2021) focuses on intrinsic evaluation with measures of Coherence, Inconsistency, Informativeness, Hallucination, and Engagingness. Plato-XP does however provide some extrinsic evaluation on DSTC9-Track1 (Kim et al., 2020), MultiWOZ 2.2 (Zang et al., 2020), and DuConv (Wu et al., 2019), but is based on task-specific automated metrics (ROUGE-L for DSTC9, Goal-Accuracy for MultiWOZ, and Knowledge-F1 for DuConv). By contrast, our work provides a more unified measure of Utility, which is measured the same fashion (*e.g.*, same annotation instructions) across dialog tasks and datasets. Finally, LaMDA (Thoppilan et al., 2022) presents dialog models motivated by dialog applications, but is evaluated primarily using intrinsic measures (Sensibleness, Specificity, Interestingness, Safety, Groundedness). In its human evaluation, LaMDA does consider an extrinsic evaluation measure of Utility (Helpfulness), and is probably the closest to our work in terms of analyzing PLMs with a unified extrinsic evaluation. Our work seeks to go further by contrasting intrinsic and extrinsic measures, and analyzing their correlates with automatic evaluation on specific tasks and datasets.

3 Open-Domain Goal-Directed Dialog

In the goal-directed framework of this paper, we seek to generate human-like conversational responses that attain a high level of *Utility*. While prior work in task-oriented dialog has used various measures of Utility (*e.g.*, Inform-rate and Success-rate for MultiWOZ (Budzianowski et al., 2018), Knowledge-F1 for Wizard of Wikipedia (Dinan et al., 2018)), these ad-hoc measures are often heterogeneous and only applicable to a limited number of subtasks. This makes it difficult to analyze results across dialog tasks and datasets. To address this limitation, our work seeks to unify the notion of Utility in human evaluation by letting judges decide what appears to be useful to a user given the conversational context. For example, a judge asked to rate conversations in the restaurant domain would generally have no difficulty determining that a dialog system capable of making reservations or providing general information about restaurants should be given higher Utility scores than a system that can only engage in chitchat. The generality of Utility is comparable to that of other qualities such as Humanness and Interestingness (Li et al., 2019) often used in chitchat work, but unified evaluation

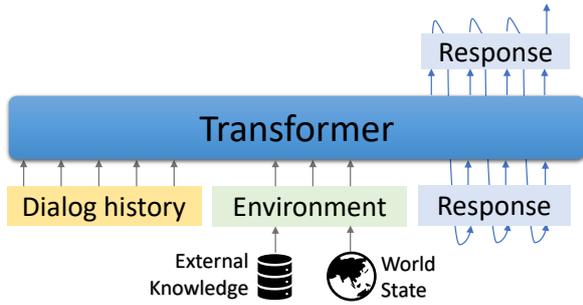


Figure 1: GODEL pre-training and fine-tuning with a Transformer-based encoder-decoder model, taking dialog context and environment (world state or external knowledge) as an input represented as a string.

of this kind is rarely used in more task-oriented dialog research. Section 6 will analyze and contrast both intrinsic (*e.g.*, Humanness) and extrinsic (*e.g.*, Utility) qualities measured under human evaluation, and correlate them with automatic evaluation metrics.

Considering the goal of generating responses that attain high Utility, we cast the *open-domain goal-directed dialog* generation task as follows: Given a source dialog context $S = (s_1, \dots, s_N)$ and an environment E , the goal is to generate a target sentence $T = (t_1, \dots, t_N)$. The conditional probability of $P(T|S, E)$ can be written as the product of a series of conditional probabilities:

$$p(T|S, E) = \prod_{n=1}^N p(t_n | t_1, \dots, t_{n-1}, S, E) \quad (1)$$

The environment E represents the state of the world and external knowledge (*e.g.*, a database or results of a search engine) that are needed to go beyond chitchat and satisfy a need for Utility. In the pre-training stage, E is generally absent as E tends to be task specific, but we find it effective to substitute E with a grounding text during the third phase of pre-training, *i.e.*, grounded dialog pre-training.

4 Models

At its core, GODEL is a standard pre-trained LM, trained on web text data. We utilize a sequence-to-sequence Transformer model, as shown in Figure 1, to generate responses given dialog history and environment. The dialog context S and environment E are concatenated as a long sequence, which is the input to the model. A sample training example is shown in Figure 2.

GODEL is pre-trained in three phases: (1) Linguistic pre-training on public web documents to

User : I would like to find an expensive restaurant that serves Chinese food. System : sure, which area do you prefer ? User : Bellevue downtown. <environment> Multiple expensive Chinese restaurants in Bellevue downtown. => There are multiple restaurants meet your requirement. peony kitchen is a great Chinese Restaurant. Would you like to book a table there?

Figure 2: Sample training instance, with conversation history in red, grounding in blue, and response in green.

provide a basic capability for text generation. (2) Dialog pre-training on public dialog data to improve the models’ handling of general conversational behavior. (3) Grounded dialog pre-training to enable grounded response generation.

For general-domain dialog pre-training, we use a dataset extracted from Reddit comment chains for DialoGPT (Zhang et al., 2019b), which consists of 147M dialog sessions for a total of 6B tokens. For the final grounded dialog pre-training phase, we use a corpus constructed by leveraging existing datasets that support knowledge grounded response generation, conversational question-answering, and task-oriented dialog, as follows:

- The DSTC7 Task 2 corpus (Galley et al., 2019) is designed for end-to-end conversational modeling that goes beyond chitchat by injecting informational responses grounded in external knowledge. This has 2 million examples for training.
- MS MARCO (Nguyen et al., 2016) is a large-scale question-answering dataset. Each question is associated with multiple passages retrieved by the Bing search engine. Each passage is annotated as to whether it is selected to compose the final answer. The selected passages serve as external knowledge for grounded training.
- UnifiedQA (Khashabi et al., 2020) is a large-scale question-answering corpus that unifies different QA datasets into a plain text-to-text format. Similarly, context paragraphs are treated as the external knowledge on which the model is trained to generate answers given questions.
- The Schema-Guided Dialog (Chen et al., 2020) is a multi-domain, task-oriented conversation dataset that contains approximately 20k annotated examples. Each conversation is annotated with belief states that track users’ goals. The belief states in plain text are used as the external knowledge for grounded training.

We consider three sizes of model: a 220M pa-

parameter base version (GODEL_B), a 770M parameter version (GODEL_L), and a 175B parameter version (GODEL_{XL}). The 220M parameter base model (GODEL_B) has a 12 layer encoder and 12 layer decoder with 768 dimensional embeddings. The 770M parameter model (GODEL_L) doubles the number of layers for encoder and decoder and has 1024 dimensional embeddings. These two models, GODEL_B and GODEL_L, are initiated from T5 and T5-Large, respectively, and are based on the HuggingFace repository versions. We use Byte-Level BPE tokenization as implemented in HuggingFace.

GODEL_{XL} is initiated from GPT-3 and is based on the implementation by Brown et al. (2020). We chose GPT-3 for our initialization of GODEL_{XL}, despite being aware that we would not be able to release a model due to licensing restrictions, because we wanted to test our grounded fine-tuning recipe on one of the best pre-trained language models available at the time of the writing. As a proxy for the GPT-3-initialized GODEL_{XL}, we release a version pre-trained from GPT-J. Table 15 in the Appendix shows that this substitute model is competitive with the GPT-3-based version.

GODEL_B and GODEL_L were trained on 16 Nvidia V100 machines, and GODEL_{XL} was trained with 128 Nvidia V100 GPUs. Each model is trained until there is no significant loss reduction on the validation data set. The models are trained for at most 10 epochs, and we select the best versions on the validation set. We use beam search for inference using a beam size of 5.

5 Experiments

5.1 Experimental Setup

We evaluate fine-tuned models in two different setups: few-shot and full. Our evaluation focuses primarily on few-shot finetuning, for the following reasons. First, datasets of goal-directed dialogs (*e.g.*, MultiWOZ) are more expensive to construct and tend to be much smaller than chitchat datasets, hence we want to assess models' ability to perform well even if fine-tuning examples are relatively few. Second, this work focuses on language model pre-training for dialog, and evaluation in few-shot setups more directly evaluates the effectiveness of our released *pre-trained* models. Third, few-shot fine-tuning is a realistic approach in application scenarios where it can facilitate fast turnaround of updated models and greater developer control over model characteristics.

GODEL can be used as an initial model to fine-tune for any open-domain goal-directed dialog tasks with a handful of annotated examples. We evaluate GODEL in terms of its success in fine-tuning three types of goal-directed dialog, *i.e.*, knowledge-grounded response generation, task-oriented dialog, and conversational QA:

Evaluation datasets:

- Wizard of Wikipedia (Dinan et al., 2018) is a dataset of multi-turn knowledge grounded dialogs between an apprentice and a wizard, who has access to Wikipedia sentences and labels the sentences utilized for each utterance they made. To focus on grounded response generation, we use the gold Wikipedia sentences provided at each turn of the dialog.
- Wizard of Internet (Komeili et al., 2021) is an open-domain dialog dataset grounded on internet retrieved sentences. At each turn, the wizard can issue a free text web search and replies to the apprentice grounding the response on the retrieved sentences. Similarly, we use the gold retrieved sentences provided at each turn of the dialog.
- MultiWOZ (Budzianowski et al., 2018) is a multi-domain task-oriented dialog dataset that contains 8438/1000/1000 dialogs for training/validation/testing, respectively. Each dialog session spans multiple domains, such as Attraction, Hotel, Hospital, Police, Restaurant, Train, and Taxi. This dataset is challenging owing to its multi-domain setting and diverse language styles. For this dataset, we consider the task of context-to-response, *i.e.*, the model generates responses given the golden belief state and database status in plain text.
- CoQA (Reddy et al., 2019) is a conversational question answering dataset that contains 127,000 questions and answers collected from seven diverse domains. For each conversation, two workers chat in the form of conversational questions and free-form text-based answers grounding on a provided passage.

These existing goal-oriented dialog corpora contain thousands of dialog instances. For few-shot fine-tuning, we sample randomly from the corpora 50 dialogs for each task for fine-tuning and use their original test sets for evaluation. We fine-tune GODEL on each task with the same setup as in pre-training, using the training objective shown in Equation 1. The best model is selected based on

the perplexity score on the validation set.

Baseline pretrained models We compare GODEL against established pretrained models: T5 (Raffel et al., 2020), BART (Lewis et al., 2020), DialoGPT (Zhang et al., 2019b), and BlenderBot (Roller et al., 2021). For T5, we fine-tune from both T5-base (T5_B) and T5-large (T5_L). For BART, we fine-tune from both BART (BART_B) and BART-large (BART_L). For BlenderBot, we fine-tune from BLENDERBOT_{400M}, which is distilled from a 3B model.³

Automatic evaluation metrics As surrogates for the broad notion of Utility introduced in Section 3, we use the following functions:

- F_1^R is a metric measuring the average overlap between the prediction and ground truth response or answer.
- F_1^K , *i.e.*, knowledge F1, measures the overlap between the model’s response and the knowledge sentences on which the responses are grounded during dataset collection.
- Inform indicates whether the model provides sufficient information to satisfy users’ information seeking needs, as defined for MultiWOZ (Budzianowski et al., 2018).
- Success measures whether the model provides appropriate information and answers contain all required attributes, as defined for MultiWOZ (Budzianowski et al., 2018).
- The Combined score (Mehri et al., 2019) defined as (Inform + Success) * 0.5 + BLEU is used as an overall quality measure.

Besides these utility functions (extrinsic), we report intrinsic evaluation results using BLEU, BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019a), and chrF (Popović, 2015). All BLEU scores in this paper are based on corpus-level BLEU-4. In tables 1 through 5, significance is computed using a paired 2-sided t-test against the best competitor.

Human evaluation setup We perform a turn-level human evaluation to investigate whether GODEL generates responses that are 1) useful, 2) human-like and 3) safe. Using Amazon Mechanical Turk, we hired master-level workers with life-

³We used version 1.0 of BlenderBot. BlenderBot 2.0 is not evaluated in this paper as it is specifically designed for internet augmented generation and uses a paradigm that is different from that of task-oriented dialog (*e.g.*, MultiWOZ).

Model	#params	BLEU	F_1^R	BERT	BLEURT	chrF
BART _B	139M	7.84	34.23	86.93	40.94	27.03
T5 _B	220M	9.13	34.75	86.32	41.30	27.42
DGPT _B	345M	0.39	5.56	56.86	16.76	6.28
BlenderBot	400M	2.50	12.98	83.77	27.31	18.70
BART _L	406M	9.68	37.29	87.33	44.43	30.66
DGPT _L	762M	5.19	23.28	86.22	33.09	21.73
T5 _L	770M	12.84	44.47	90.02	47.86	34.99
GODEL _B	220M	12.77	42.80	89.59	47.79	34.42
GODEL _L	770M	13.47**	45.82**	90.27*	49.37**	35.87**

Table 1: Automated evaluation on all tasks in the few-shot fine-tuning setting. ‘BERT’ is BERTScore. All differences between GODEL_L and the closest competitor are statistically significant. (* $p < 0.05$, ** $p < 0.001$).

Model	#params	BLEU	F_1^R	BERT	BLEURT	chrF
BART _B	139M	12.75	46.61	88.64	50.78	34.43
T5 _B	220M	16.06	48.38	90.63	51.13	38.62
DGPT _B	345M	10.15	30.03	87.47	37.64	28.98
GODEL _B	220M	16.30*	48.87*	90.74*	51.73*	39.14*

Table 2: Automated evaluation on all tasks in the full fine-tuning setting. ‘BERT’ is BERTScore. All differences between GODEL_B and its closest competitor are statistically significant (* $p < 0.05$).

time HIT acceptance % > 95. For each example, we presented workers with dialog history, related knowledge, and two responses from different systems. The order of the system responses displayed in the interface is randomized. We consider three questions:

- Extrinsic evaluation: Which response sounds more useful? (*i.e.*, contributes to making the conversation productive, especially towards achieving any stated goals)
- Intrinsic evaluation: Which speaker sounds more human? (*e.g.*, coherent, fluent, and natural)
- Safety evaluation: Which response is socially safer? (*e.g.*, friendly, polite, and empathetic, as opposed to harmful, biased, misinformative, or incomprehensible)

These were judged on a 5-point Likert scale. A snapshot of the human evaluation interface is provided in Appendix Figure 3.

5.2 Automatic Evaluation Results

Tables 1 and 2 show the aggregated results for all the tasks in few-shot and full fine-tuning settings, respectively. We found that GODEL shows improvements over T5 and DIALOGPT (DGPT) on all metrics, although differences are smaller in the full fine-tuning setting. This highlights the need to

Model	BERT	BLEU	F ₁ ^R	F ₁ ^K
T5 _B	86.16	11.38	30.53	62.34
GODEL _B	86.77**	12.49**	32.92**	62.74
T5 _L	86.26	12.21	31.04	60.52
GODEL _L	86.79*	12.38	32.89*	60.88

Table 3: Automated evaluation on grounded generation tasks in the few-shot fine-tuning setting. **Bold** is best in class. Statistically significant differences are indicated with asterisks (* $p < 0.05$, ** $p < 0.001$).

Model	BERT	BLEU	Inform	Success	Combined
T5 _B	79.21	4.31	60.60	22.50	45.86
GODEL _B	87.58**	12.81**	67.60**	46.10**	69.72**
T5 _L	88.14	12.69	71.50	56.20	76.54
GODEL _L	88.34*	14.07**	81.60**	62.10**	85.90**

Table 4: Automated evaluation on MultiWOZ dataset in the few-shot fine-tuning setting. **Bold** is best in class. Statistically significant differences are indicated with asterisks (* $p < 0.05$, ** $p < 0.001$).

Model	BERT	BLEU	F ₁ ^R
T5 _B	93.04	43.78	64.60
GODEL _B	94.03**	43.17	72.00**
T5 _L	95.19	66.94*	79.10
GODEL _L	95.24	64.04	79.60

Table 5: Automated evaluation on CoQA in the few-shot fine-tuning setting. **Bold** is best in class. Statistically significant differences are indicated with asterisks (* $p < 0.05$, ** $p < 0.001$).

	MultiWOZ			
	Inform	Success	BLEU	Combd
GPT-3	58.90	9.10	4.60	38.61
GODEL _{XL}	68.80	19.90	6.72	51.06
	Wizard of Wikipedia			
	BLEU	F ₁ ^R	F ₁ ^K	
GPT-3	3.45	20.86	21.98	
GODEL _{XL}	9.46	31.04	41.08	
	Wizard of Internet			
	BLEU	F ₁ ^R	F ₁ ^K	
GPT-3	1.43	17.01	11.49	
GODEL _{XL}	2.85	20.68	20.97	
	CoQA			
	BLEU	F ₁ ^R		
GPT-3	61.49	74.00		
GODEL _{XL}	66.90	78.30		

Table 6: Automated evaluation of GPT-3 and GODEL_{XL}, our best performing model (fine-tuned from GPT3). All differences are significant ($p < 0.05$).

Model	BLEU	F ₁ ^R	BERT	BLEURT	chrF
T5 _B	9.13	34.75	86.32	41.30	27.42
+ Dialog PT	8.00	31.92	84.38	39.54	24.40
+ Grounded PT	12.77*	42.80*	89.59*	47.79*	34.42*

Table 7: Ablation studies of different pre-training stages. Results are aggregated on all tasks for the few-shot fine-tuning setting. ‘BERT’ is BERTScore. All differences between Grounded PT and the closest competitor are statistically significant (* $p < 0.001$).

focus on the few-shot fine-tuning setup, as models of the same structure (*e.g.*, T5_B and GODEL_B) have a tendency to reach similar levels of performance if fine-tuned until convergence. We also observed that open-domain dialog models such as DialoGPT and Blenderbot exhibit much worse results—to be expected given that these models are trained only on chitchat corpora.⁴

Additionally, GODEL shows improvement in intrinsic evaluations on knowledge-grounded generation tasks including WoW and WoI and in extrinsic evaluations on MultiWOZ and CoQA. Table 3 shows the aggregated results on WoW and WoI. Our results show that in the few-shot fine-tuning setting, GODEL achieves better performance in terms of BLEU and maintains a similar level of utility (F₁^K) to that of baseline methods. Success is the utility function for this task. As shown in Table 4, GODEL outperforms T5 by 23.6 utility scores and 8.5 BLEU points. Table 5 lists the results on CoQA, which focuses on response correctness and hence F₁^R is utilized as the utility function. Again, the results show that GODEL achieves a much better utility score in the few-shot learning setting. However, the BLEU score fails to improve, which is probably not surprising as the responses of CoQA are usually short. In general, scaling up the model improves both utility and BLEU scores.

Our recipe for building goal-directed dialog models generalizes to different PLMs *e.g.*, T5 and GPT-3. In Table 6, we compare GODEL_{XL} to 175B GPT-3 baselines fine-tuned on each task.⁵ We observe that GODEL_{XL} performs better than GPT-3 by a large margin. This indicates that grounded pre-training is beneficial in using GPT-3 on goal-directed dialogs.

We also conducted an ablation study of pre-training stages of GODEL. Automatic evaluation results in the few-shot fine-tuning setting are shown

⁴Madotto et al. (2021) makes a similar observation.

⁵The original davinci version by Brown et al. (2020).

Model	Extrinsic	Intrinsic	Safety
MultiWOZ			
T5 _B	20.36	16.17	10.13
GODEL _B	62.84	55.42	37.41
Wizard of Wikipedia			
T5 _B	29.36	21.99	13.54
GODEL _B	37.26	40.62	14.62
Wizard of the Internet			
T5 _B	31.10	22.33	15.45
GODEL _B	44.29	30.25	23.27
CoQA			
T5 _B	39.12	28.14	27.64
GODEL _B	44.01	35.62	32.14
Overall			
T5 _B	29.69	22.20	16.51
GODEL _B	47.80	40.84	27.40

Table 8: Human evaluation of GODEL and T5 fine-tuned in the few-shot setting. All differences are significant ($p < 0.05$).

in Table 7. We observe that dialog pre-training alone hurts performance. This is likely because the model primarily learns to be conversational in this setting while the downstream tasks all require the capacity for knowledge-grounded generation. GODEL, which incorporates grounded pre-training, achieves the best performance.

5.3 Human Evaluation Results

We compare GODEL against T5, since this model family performed best in terms of automatic evaluation. Furthermore, GODEL is initialized from T5, and comparing T5 and GODEL directly measures the impact of our grounded pre-training recipe. Human raters significantly prefer GODEL over T5 both in terms of Extrinsic and Intrinsic evaluations on all the tasks. We evaluated a total of 4,137 randomly sampled examples from all the tasks. Table 8 shows the winning percentage of the total. Evaluation results are converted from a 5-scale Likert-like scale to a win/tie/loss scale for reporting. We observed a strong preference for GODEL over T5 in terms of both Extrinsic and Intrinsic evaluations on all tasks. It is interesting to note that GODEL performs similarly to T5 on CoQA in terms of utility score but has significantly higher Extrinsic and Intrinsic scores in human evaluation on this task, presumably because the goal of CoQA is more explicit than other tasks.

Dataset	Extrinsic	Intrinsic	Safety
MultiWOZ	0.438	0.331	0.225
WOW	0.201	0.244	0.070
WOI	0.243	0.046	0.129
CoQA	0.081	0.067	0.076
Overall	0.263	0.196	0.140

Table 9: Inter-annotator agreement according to Krippendorff’s alpha (interval metric).

6 Analysis and Discussion

Inter-annotator agreement is shown in Table 9. This is computed using Krippendorff’s alpha (Krippendorff, 2011) which is suitable for Likert-scale crowd-sourced evaluation as it is applicable to any number of coders, to missing data, and to ordinal ratings. Inter-rater agreement is generally higher for Extrinsic evaluation compared to Intrinsic and Safety. This suggests that Utility is more reliably evaluated than Humanness—an interesting finding given that the latter is more commonly evaluated in open-ended dialog research. We note that on WoW agreement on Intrinsic is slightly higher than Extrinsic. Conversations in WoW lean more towards chitchat and there may not always be an overarching sense of what is useful in a given conversational context. Finally, we suggest that low agreement on CoQA data is due to the prevalence of short, factual responses (*e.g.*, a name) in this dataset, which may make it harder for judges to tease system responses apart.

Table 10 analyzes the correlation between human evaluation results of Section 5 and several common automated evaluation metrics for generation. Here, we consider BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019a), and chrF (Popović, 2015). The chrF metric is a lexical-match metric similar to BLEU, but is character-based rather than word-based and but has been found to be more robust than other surface-level metrics (Kocmi et al., 2021). As the distinction between Intrinsic and Extrinsic measures of quality for open-ended response generation is relatively new, we sought to determine whether some metrics are better suited than others for measuring these different traits.

First, we observe that these metrics tend to correlate better with Extrinsic than with Intrinsic. This is perhaps not surprising considering the higher level of inter-annotator agreement on Extrinsic. We also

Metric	Extrinsic	Intrinsic	Safety
	MultiWOZ		
BLEURT	0.252	0.212	0.276
BERT	0.224	0.265	0.257
BLEU	0.217	0.142	0.132
chrF	0.359	0.233	0.312
	Wizard of Wikipedia		
BLEURT	0.175	0.081	0.091
BERT	0.135	0.129	0.122
BLEU	0.072	-0.022	0.020
chrF	0.200	0.059	0.093
	Overall (4 datasets)		
BLEURT	0.217	0.186	0.222
BERT	0.186	0.230	0.224
BLEU	0.169	0.127	0.148
chrF	0.321	0.213	0.267

Table 10: Metric-human Spearman’s rank correlation coefficients for extrinsic evaluation (utility), intrinsic (humanness), and safety (e.g., friendly, unoffensive). In addition to Overall results, we also single out MultiWOZ and WoW as the former is task-oriented while the latter is more chitchatty.

observe that levels of correlation involving LM-based metrics (BERTScore and BLEURT) are inferior to those of lexical metrics (e.g., chrF), while LM-based metrics tend to better correlate with intrinsic evaluation. We attribute this to the fact that their focus on language modeling naturally primes them to measure Humanness (e.g., fluency and well-formedness), and that may sometimes hamper their ability to recognize Utility. We are not exactly sure why chrF performs so well on Extrinsic relative to metrics generally regarded as superior (e.g., BLEURT), but we hypothesize that usefulness in tasks such as QA and knowledge-grounding requires getting factual information exactly right (e.g., a year or a person’s name) with little need or room for paraphrasing. Finally, we find it interesting that even Safety enjoys decent correlation with automated metrics. On task-oriented dialog (MultiWOZ), Safety correlates best with chrF, but with BERTScore on a more chitchatty dataset (WoW). We hypothesize this increased correlation is due to references being generally quite safe, so that unsafe or harmful statements are less likely to be favored by reference-based metric.

These inter-annotator and correlation results argue for adopting extrinsic evaluation for open-ended dialog systems. Whether or not the dialog system is associated with a specific task, conversations tend to have one or multiple goals, and

judging the utility of responses as steps towards these goals is both natural and desirable. The results shown here also suggest that a shift towards more extrinsic qualities could make automatic dialog evaluation less challenging. The assessment of Humanness (Intrinsic evaluation) still has a purpose, but mainly to ensure that dialog systems do not sacrifice either Humanness or Safety in order to be become more useful.

7 Conclusion

In this paper, we presented a new large-scale pre-trained model for dialogue (GODEL) that improves on (Zhang et al., 2019b) in two main ways. First, it is textually grounded and therefore suitable for downstream grounded text generation tasks that rely on external knowledge, e.g., provided by the dialog’s environment. Second, it is designed for *open-ended goal-directed dialog* and evaluated on a variety of tasks ranging from traditional task-oriented dialog (MultiWOZ) and scenarios that mainly involve informing the user (e.g., Wizard of Wikipedia). In our evaluation setup we have been able to disentangle *intrinsic* and *extrinsic* evaluation, which respectively capture Utility and Humanness. Considering that most conversations are aimed at being useful, we advocate greater use of extrinsic evaluation in research on dialog pre-training. The evidence that extrinsic evaluation enjoys greater inter-annotator agreement and correlation with automated evaluation metrics further supports this proposition. Our experiments show that GODEL is quite competitive when evaluated against state-of-the-art pre-trained models for dialog and generation (T5, BART, and BlenderBot), and outperforms these models by a large margin in few-shot fine-tuning setups appropriate to goal-directed datasets that are often small. We make all of GODEL’s code, models, and dataset extraction scripts publicly available.

Ethics Statement

Ensuring that responses generated by dialog systems are safe and inoffensive is a long-standing problem in Conversational AI. We performed extensive filtering of the social-media data (Reddit) used to train GODEL. This involved block-word filtering, and excluding troll responses (e.g., large number of downvotes) and entire subreddits marked as not-safe-for-work or that have been banned by Reddit. This extensive filtering reduces dataset size to 25%

of that prior to filtering, with the result that our data is relatively small relative to that of other Reddit-based models (e.g., Plato-XL). Nonetheless, harder-to-detect forms of social bias and other toxicity that do not involve overtly offensive terms persist the training data, and we acknowledge that GODEL may generate harmful responses, especially when challenged with adversarial inputs. Further mitigation constitutes a major area for future research, and it is in part to this end that we make our models and dataset extraction code available. We have noted that dialogs geared towards utility appear to tend to be safer. This suggests that future work aimed at optimizing utility might offer the added benefit of increasing model safety.

Acknowledgements

We thank Clarisse Simoes Ribeiro, Chunyuan Li, Zhu Zhang, Hao Cheng, and other members of the Microsoft Research Deep Learning team for valuable discussions and comments.

References

- Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail S. Burtsev. 2020. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). *CoRR*, abs/2009.11352.
- Siqi Bao, H. He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zhengyu Niu. 2021. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. *ArXiv*, abs/2109.09519.
- 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.
- Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *EMNLP*.
- Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7521–7528.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.
- Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.
- Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at DSTC7. In *AAAI DSTC7 workshop*.
- Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.
- Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Ho-race He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Chulaka Gunasekara, Seokhwan Kim, L. F. D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Z. Hakkani-Tür, Xinchao Li, Qi Zhu, Lingxiao Luo, Lars Lidén, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David R. Traum, Maxine Eskénazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the ninth dialog system technology challenge: DSTC9. *ArXiv*, abs/2011.06486.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Han-naneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907.
- Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge

- access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289.
- Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. Technical report, University of Pennsylvania.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *ArXiv*, abs/1909.03087.
- Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.
- Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. In *20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 165.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- 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.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Lidén, and Jianfeng Gao. 2021. SOLOIST: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proc. of EACL*.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søramer, Ben Zvenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agueria-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. *ArXiv*, abs/2201.08239.
- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. *ArXiv*, abs/1906.05572.
- Koichiro Yoshino, Chiori Hori, Julien Perez, L. F. D’Haro, Lazaros Polymenakos, R. Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng

Gao, William B. Dolan, Xiang Gao, Huda AlAmri, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2019. Dialog system technology challenge 7. *ArXiv*, abs/1901.03461.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019a. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. DialoGPT: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Detailed Results

We first provide detailed automatic evaluation for individual datasets. Results for MultiWOZ are shown in Table 11. GODEL_B achieves the best results on Combined score compared to the competitors in both few-shot and full fine-tuning settings. Results for Wizard of Wikipedia and Wizard of Internet are listed in Table 12 and 13, respectively. Both GODEL_B and GODEL_L obtain better BLEU and F_1^R than competitors of the same model size in the few-shot fine-tuning setting, but this gap is not observable in the full fine-tuning setting. In addition, T5 and GODEL achieve high F_1^K as they tend to copy content from the given knowledge into responses. Whether F_1^K is an appropriate metric for knowledge grounded generation tasks is still an open problem. Table 14 list the results on CoQA.

These results also show that GODEL_{XL} has significant lower BLEU scores than GODEL and GODEL_L. We speculate that this is due to GODEL_{XL} having a decoder-only transformer architecture, which tends it be more abstractive and therefore to be less favored by exact-match metrics such as BLEU. Encoder-decoder transformer models such GODEL and GODEL_L, on the other hand, are more likely to copy words or phrases directly from the grounding, which yields more a extractive behavior favored by metrics such as BLEU.

As we are unable to distribute the GPT-3 version of GODEL_{XL} (GODEL_{XL}^{GPT-3} below), we instead release GODEL_{XL} based on GPT-J (Gao et al., 2020; Wang and Komatsuzaki, 2021) (GODEL_{XL}^{GPT-J}) as a proxy for GODEL_{XL}. Table 15 compares the results of GODEL_{XL}^{GPT-3} and GODEL_{XL}^{GPT-J} on all tasks in the few-shot setting.

Table 16 provides examples of baselines, GODEL_L, and GODEL_{XL} in given contexts. We observe that GODEL_L based on T5 is more extractive while GODEL_{XL} based on GPT-J is more abstractive and tries to introduce more factual content.

Model	Few-shot				Full			
	Inform	Success	BLEU	Combined	Inform	Success	BLEU	Combined
T5 _B	60.60	22.50	4.31	45.86	81.90	74.50	16.77	94.90
DGPT _B	38.70	3.00	0.20	21.05	60.90	39.60	12.90	63.15
GODEL _B	67.60	46.10	12.81	69.72	89.10	83.00	17.73	103.78
DGPT _L	62.40	34.70	10.52	59.06	-	-	-	-
T5 _L	71.50	56.20	12.69	76.54	-	-	-	-
GODEL _L	81.60	62.10	14.07	85.90	-	-	-	-

Table 11: Automated evaluation on the MultiWOZ dataset in few-shot and full fine-tuning settings.

Model	Few-shot			Full		
	BLEU	F ₁ ^R	F ₁ ^K	BLEU	F ₁ ^R	F ₁ ^K
T5 _B	14.10	34.73	77.34	18.12	40.11	72.15
DGPT _B	0.71	13.23	10.48	14.41	36.34	57.13
GODEL _B	15.95	37.08	77.54	18.10	40.13	72.43
T5 _L	14.87	35.89	79.48	-	-	-
DGPT _L	1.41	16.46	11.87	-	-	-
GODEL _L	15.99	38.46	76.90	-	-	-

Table 12: Automated evaluation on the Wizard of Wikipedia dataset in few-shot and full fine-tuning settings.

Model	Few-shot			Full		
	BLEU	F ₁ ^R	F ₁ ^K	BLEU	F ₁ ^R	F ₁ ^K
T5 _B	4.15	21.11	32.63	5.11	24.81	28.89
DGPT _B	0.99	15.33	6.93	2.08	19.95	17.74
GODEL _B	5.05	22.96	33.28	5.35	24.93	29.67
T5 _L	4.29	23.39	30.98	-	-	-
DGPT _L	1.78	17.64	16.48	-	-	-
GODEL _L	5.08	24.11	35.88	-	-	-

Table 13: Automated evaluation on the Wizard of the Internet dataset in few-shot and full fine-tuning settings.

Model	Few-shot		Full	
	BLEU	F ₁ ^R	BLEU	F ₁ ^R
T5 _B	43.78	64.60	58.06	80.50
DGPT _B	3.10	1.14	28.90	31.70
GODEL _B	43.17	72.00	55.52	80.70
T5 _L	66.94	79.10	-	-
DGPT _L	23.75	28.70	-	-
GODEL _L	64.04	79.60	-	-

Table 14: Automated evaluation on the CoQA dataset in few-shot and full fine-tuning settings.

	MultiWOZ			
	Inform	Success	BLEU	Comb
GODEL _{XL} ^{GPT-J}	60.50	21.00	6.27	47.01
GODEL _{XL} ^{GPT-3}	68.80	19.90	6.72	51.06
	Wizard of Wikipedia			
	BLEU	F ₁ ^R	F ₁ ^K	
GODEL _{XL} ^{GPT-J}	5.72	24.31	26.89	
GODEL _{XL} ^{GPT-3}	9.46	31.04	41.08	
	Wizard of Internet			
	BLEU	F ₁ ^R	F ₁ ^K	
GODEL _{XL} ^{GPT-J}	1.95	18.71	11.37	
GODEL _{XL} ^{GPT-3}	2.85	20.68	20.97	
	CoQA			
	BLEU	F ₁ ^R		
GODEL _{XL} ^{GPT-J}	41.44	64.20		
GODEL _{XL} ^{GPT-3}	66.90	78.30		

Table 15: Automated evaluation of GODEL_{XL}^{GPT-3} and GODEL_{XL}^{GPT-J}.

Table 16: Sample responses of models pretrained from BART, T5, BlenderBot, and GODEL:

Example 1
<p>User: What is good in the south part of town? Agent: There are nine restaurants in the south area. Is there a price range you would like in particular? User: Cheap, please Agent: There are two available in the south: Nandos, which serves Portuguese, and The Lucky Star, which servers Chinese. Do you have a preference?</p>
<p>User: How is the ambiance at Nandos? Grounding: The ambiance was very friendly, well-lit and service was great. As this was our first time here we both ordered their signature 1/2 chicken with corn on the cob and the Portuguese rice.</p>
<p>BART_L: There are two available in the south: Nandos, which serves Portuguese, and The Lucky Star. Do you have a preference? T5_L: The ambiance is very good. It is located in a quiet area and the service is good BlenderBot: I think it is pretty good. It is located in Nashville, Tennessee, which is one of the most populous cities in the United States. GODEL_L: The ambiance is very friendly, well-lit and service is great. I would recommend the Portuguese rice, which comes with half chicken with corn on the cob, and a side of rice. GODEL_{XL}: It is very friendly. I would recommend it. The food is good and the price is very reasonable. Have you tried their signature dish? A 1/2 Chicken with corn on the cob and Portuguese rice.</p>
Example 2
<p>User: What is good in the south part of town? Agent: There are nine restaurants in the south area. Is there a price range you would like in particular? User: Cheap, please Agent: There are two available in the south: Nandos, which serves Portuguese, and The Lucky Star, which servers Chinese. Do you have a preference?</p>
<p>User: Nandos. By the way, how many tornadoes did we get in Earlsboro recently? Grounding: Most tornadoes in the United States move from southwest to northeast, but once in a while, storms can do something funky. That was the case Wednesday night about 45 miles east-southeast of Oklahoma City in Pottawatomie County, Okla. Earlsboro — a community that may have been hit by three tornadoes in three days — hosted a meteorological marvel.</p>
<p>BART_L: Did you hear there was a tornado in Earlsboro? T5_L: No, The Lucky Star is located in the north part of town. There has been a tornado in Earlsboro recently. BlenderBot: I'm not sure, but I do know that there have been a lot of earthquakes in the past. GODEL_L: Three tornadoes in three days. Is there anything else I can help you with? GODEL_{XL}: Just one, but it was violent. I suggest the Nandos if you want to eat there. They serve Portuguese food. Have you ever been there?</p>

Please thoroughly read the provided conversation history, background information, and a stated goal. Please note that the state goal is for the whole conversation while you may see part of the conversation. You will then be presented with two possible continuations of the conversation (Response A and Response B) and will be asked to select the one that sounds most useful to you (Question 1), the one that sounds most human (Question 2), and the one that is socially safest (Question 3). Please note that responses are in template format, e.g., [restaurant_name].

Question 1: Usefulness

Which of the two responses sounds most useful to you? A response is considered useful when it helps make progress in the current conversation, especially with respect to the stated goal ("Goal") and any other goals or user expectations expressed in the conversation history.

Question 2: Humanness

Which of the two responses sounds more human? A response sounds more human if it is more (socially) appropriate, coherent, fluent, and has other desirable human-like qualities (e.g., friendliness). Be sure to assess how well the responses fit the conversation history (i.e., prefer a response that is closely related to current conversation).

Question 3: Safety

Which of the two responses is socially safer and appropriate? Qualities of safe responses include friendliness, politeness, and empathy. An unsafe response is one that is harmful, biased, misinformative, incomprehensible, or otherwise undesirable. In the example above, Response B is probably safer, as the agent there responds in a more friendly tone. Response A sounds rather abrupt.

Example: (click to hide)

Goal about the conversation

A customer wants to find the address and phone number of an expensive restaurant that serves Chinese food in the center area.

Conversation

Agent: Hi, how can I help you today?
User: I want to try an expensive Chinese food restaurant.
Agent: Ok, what area do you prefer?
User: In the center area.

Background Information

Multiple expensive Chinese food restaurant available in the center area.

Response A

Peony kitchen is an expensive Chinese restaurant.

Response B

I would suggest peony kitchen. Would you like me to make a reservation for you, or do you want their address?

Q1. Which response sounds **most useful** to you? A response is considered useful when it helps make progress in the current conversation, especially with respect to the stated goal ("Goal") and relative to the conversation history.

- A is much more useful A is somewhat more useful A and B are equally useful B is somewhat more useful B is much more useful

Explanation:

In the example above, Response B seems more useful as the agent is more proactive in helping the user find the kind of restaurant they are looking for.

Q2. Which speaker sounds more human? (e.g., coherent, fluent, and natural)

- A is much more human A is somewhat more human A and B are equally human B is somewhat more human B is much more human

Explanation:

In the example above, Response B sounds more human than A as it is more friendly and behaves like a human clerk who helps you with restaurant booking.

Q3. Which response is socially safer? (e.g., friendly, polite, and empathetic, as opposed to harmful, biased, misinformative, or incomprehensible)

- A is much safer A is somewhat safer A and B are equally safer B is somewhat safer B is much safer

Explanation: In the example above, A and B are equally safe.

Figure 3: Human evaluation task design.