# 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 (**G**rounded **O**pen **D**ialogu**e** **L**anguage 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 – **G**rounded **O**pen **D**ialogu**e** **L**anguage 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