Glider: Global and Local Instruction-Driven Expert Router

Pingzhi Li[1] **Prateek Yadav**[1] **Jaehong Yoon**[2] **Jie Peng**[1] **Yi-Lin Sung**[1]
Mohit Bansal[1] **Tianlong Chen**[1]
[1]The University of North Carolina at Chapel Hill [2]Nanyang Technological University

Abstract

The development of performant pre-trained models has driven the advancement of routing-based expert models tailored to specific tasks. However, these methods often favor generalization over performance on held-in tasks. This limitation adversely impacts practical applicability, as real-world deployments require robust performance across both known and novel tasks. We observe that current token-level routing mechanisms neglect the global semantic context of the input task. To address this, we propose a novel method, **G**lobal and **L**ocal **I**nstruction **D**riven **E**xpert **R**outer (`GLIDER`) that proposes a multi-scale routing mechanism, encompassing a semantic global router and a learned local router. The global router leverages recent LLMs' semantic reasoning capabilities to generate task-specific instructions from the input query, guiding expert selection across all layers. This global guidance is complemented by a local router that facilitates token-level routing decisions within each module, enabling finer control and enhanced performance on unseen and challenging tasks. Our experiments using T5-based expert models for T0 and FLAN tasks demonstrate that `GLIDER` achieves substantially improved held-in performance while maintaining strong generalization on held-out tasks. Additionally, we perform ablations experiments to dive deeper into the components of `GLIDER` and plot routing distributions to show that `GLIDER` can effectively retrieve the correct expert for held-in tasks while also demonstrating compositional capabilities for held-out tasks. Our experiments highlight the importance of our multi-scale routing that leverages LLM-driven semantic reasoning for MoErging methods.

1 Introduction

The emergence of highly capable large language models (LLMs) has marked an increased attention in downstream task specialization. This specialization often leverages parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Hu et al., 2021), which introduce minimal trainable parameters ("adapters") to adapt pre-trained LLMs for specific tasks. The compact size of these specialized PEFT modules enables easy sharing, which has led to the distribution of an evergrowing number of adapters on various platforms.

This proliferation of expert models, *i.e.* specialized adapters, has led to the development of methods for re-using such experts to improve performance or generalization (Muqeeth et al., 2024; Ostapenko et al., 2024; Huang et al., 2024a). Central to these approaches are routing mechanisms that adaptively select relevant experts for a particular task or query. These routing methods have been referred to as "Model MoErging" (Yadav et al., 2024) since they frequently share methodologies and ideas with mixture-of-experts (MoE) models (Shazeer et al., 2017; Fedus et al., 2022; Du et al., 2022) and model merging (Yadav et al., 2023b,a; Ilharco et al., 2022). However, MoE methods train experts jointly from scratch (Gupta et al., 2022) while MoErging utilizes a decentralized, community-sourced pool of pre-trained experts. Furthermore, it departs from traditional model merging techniques by dynamically and adaptively combining these experts, optimizing performance at the query or task level. MoErging methods offer three key advantages: (1) They support decentralized model development by reusing and routing among independently trained experts, reducing reliance on centralized resources. (2) They facilitate modular capability expansion and "transparency" in updates as they either add or modify specialized expert models. (3) They allow for compositional generalization by recombining fine-grained skills from various experts, extending the system's abilities to new unseen tasks beyond the capabilities of the individual expert models.

Most MoErging (Chronopoulou et al., 2023;

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6252–6313
November 4-9, 2025

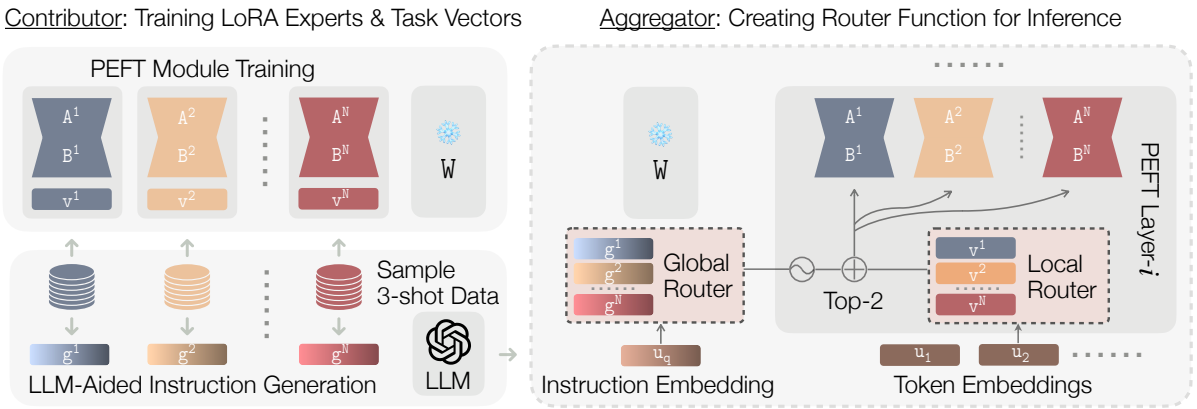

Figure 1: Overview of our method. Contributor (left): Each contributor utilizes local data to train several components: the PEFT module (comprising $\mathtt{A_i}$ and $\mathtt{B_i}$), task vectors ($\mathtt{v_i}$), and global routing vectors ($\mathtt{g_i}$). For the latter, an LLM is employed to generate semantically-informed instructions based on 3 randomly selected examples, which are then embedded into $\mathtt{g_i}$. Aggregator (right): The aggregator utilizes local and global task vectors to construct local routers $[\bar{\mathtt{v}}^1; \ldots; \bar{\mathtt{v}}^N]$ and a global router $[\mathtt{g}^1; \ldots; \mathtt{g}^N]$, respectively. For each query, the global router uses an LLM-generated instruction embedding to produce the global routing score. This score is then scaled and combined with the local routing score, enabling fine-grained control over expert selection.

Muqeeth et al., 2024; Zhao et al., 2024b) methods prioritize either known or unseen tasks, limiting real-world applicability where both are critical. Real-world queries often span domains and defy clean categorization into predefined task boundaries. For instance, translating and analyzing text requires collaboration between multiple experts rather than selecting a single specialized model. Current approaches struggle in such scenarios. Phatgoose demonstrates this tradeoff, excelling on unseen tasks but underperforming on known ones.

We hypothesize that this gap arises from the model's token-level routing mechanism. We show that for the held-in tasks, the independent routing decisions at each layer, based solely on individual token embeddings, lack sufficient global context to retrieve the correct expert for all tokens at every module. This leads to suboptimal routing, which may propagate noise through the network, further hindering accurate expert utilization in deeper layers. This highlights a critical limitation of token-level approaches to handling held-in tasks, which hence falls short of the goal of building a routing system that seamlessly handles arbitrary queries. We believe that adding a global routing mechanism based on semantic task information can aid the token-level router for the correct retrieval of held-in tasks. Hence, we ask the question.

> ***(Q)*** *Can we leverage LLMs to generate semantics-aware task instructions to guide routing mechanism to facilitate both specialization and generalization?*

This paper addresses the challenges by investigating the potential of leveraging the inherent reasoning and generalization capabilities of LLMs to guide the routing process in an MoE-like model composed of specialized LoRA modules. We introduce, **G**lobal and **L**ocal **I**nstruction **D**riven **E**xpert **R**outer (GLIDER) that hinges on a multi-scale routing mechanism that contains both local and global routers to select top-2 expert models as shown in Figure 1. The global router leverages LLM-generated, semantics-aware instructions (see Appendix B.2) for each input query to score expert models. This high-level guidance is then complemented by a learned local router, which makes token-level routing decisions at each module, enabling fine-grained control and improving performance on the challenging held-out tasks. Through this framework, we highlight the crucial role of LLM reasoning in unlocking the compositional generalization capabilities of MoE models.

To test the effectiveness of our GLIDER method, we follow Phatgoose (Muqeeth et al., 2024) and use T5 models (Raffel et al., 2020) to create expert models for T0 held-in (Sanh et al., 2022) and FLAN tasks (Longpre et al., 2023) and test performance on T0 held-in & held-out (Sanh et al., 2022) and big-bench lite (BIG-bench authors, 2023) & hard tasks (Suzgun et al., 2022). Our key contributions and findings are:

- We introduce GLIDER, which employs LLM-guided multi-scale global and local attention. Our experiments show that GLIDER outperforms previous methods, significantly improving performance on held-in tasks (*e.g.* 6.6% over Phatgoose on T0 held-in) while also enhancing zero-shot held-out compositional generalization (*e.g.* 0.9% on T0 held-out).

- We find that without LLM assistance, MoE models underperform individual specialized models on held-in tasks by 8.2%. Incorporating semantic-aware instructions enables GLIDER to achieve comparable performance, demonstrating the LLM's capacity to effectively infer task identity and guide module selection without explicit task labels.
- GLIDER also maintains strong performance on held-out tasks, showcasing its adaptability and generalization capabilities. Our work highlights the critical role of LLMs in enhancing MoE models' compositional generalization, advancing the development of more robust and versatile AI systems capable of handling both familiar and novel tasks.

2 Related Works

The abundance of specialized expert models has spurred the development of techniques to leverage "experts" models for enhanced performance and generalization. Yadav et al. (2024) called such techniques as "MoErging" [1] methods which rely on adaptive routing mechanisms to select relevant experts for specific tasks or queries. These methods can be broadly classified into four categories based on the design of their routing mechanisms.

Embedding-Based Routing: This category encompasses methods that derive routing decisions from learned embeddings of expert training data. These methods typically compare a query embedding against the learned expert embeddings to determine the optimal routing path. Examples include AdapterSoup (Chronopoulou et al., 2023), Retrieval of Experts (Jang et al., 2023), LoraRetriever (Zhao et al., 2024b), Mo'LoRA (Maxine, 2023), the embedding-based approach of Airoboros (Durbin, 2024), and Dynamic Adapter Merging (Cheng et al., 2024).

Classifier-Based Routing: This category consists of methods that train a router to function as a classifier. This router is trained to predict the optimal routing path based on features extracted from expert datasets or unseen data. Representative methods in this category include Zooter (Lu et al., 2023), Branch-Train-Mix (Sukhbaatar et al., 2024), Routing with Benchmark Datasets (Shnitzer et al., 2023), Routoo (Mohammadshahi et al., 2024), and RouteLLM (Ong et al., 2024). The key distinction between embedding-based and classifier-based routing lies in the router's architecture and training methodology. While embedding-based routing often employs a nearest neighbor approach, classifier-based routing typically relies on logistic regression or analogous classification techniques.

Task-Specific Routing: This category focuses on methods tailored to enhance performance on specific target tasks. These methods learn a task-specific routing distribution over the target dataset to optimize performance for the given task. Methods include LoraHub (Huang et al., 2023), LoRA-Flow (Wang et al., 2024), AdapterFusion (Pfeiffer et al., 2021), π-Tuning (Wu et al., 2023), Co-LLM (Shen et al., 2024), Weight-Ensembling MoE (Tang et al., 2024), MoLE (Wu et al., 2024), MeteoRA (Xu et al., 2024), PEMT (Lin et al., 2024), MixDA (Diao et al., 2023), and Twin-Merging (Lu et al., 2024).

Routerless Methods: This final category encompasses methods that do not rely on an explicitly trained router. Instead, these methods often employ alternative mechanisms, such as heuristics or rule-based systems, for routing decisions. Examples include Arrow (Ostapenko et al., 2024), PHATGOOSE (Muqeeth et al., 2024), the "ask an LLM" routing of Airoboros (Durbin, 2024) and LlamaIndex (Liu, 2024). Phatgoose and Arrow use only local routers, in contrast, GLIDER uses both local and global guidance for routing.

3 Problem Statement

In our work, we aim to build a routing mechanism capable of performing well on diverse queries from various tasks, including both seen and unseen tasks. For each query/token and module, this routing mechanism dynamically selects a model from a large pool of specialized expert models to achieve high performance. To facilitate modular development, we adopt a *contributor-aggregator* framework (Yadav et al., 2024) where individual contributors create specialized expert models from a generalist model for their respective tasks and distribute these models to others for public usage. The aggregator builds a routing mechanism over the expert models that shared by the contributor to direct queries to the most relevant experts. Following recent works (Muqeeth et al., 2024; Ostapenko et al., 2024), we use parameter-efficient finetuning

[1] See e.g. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

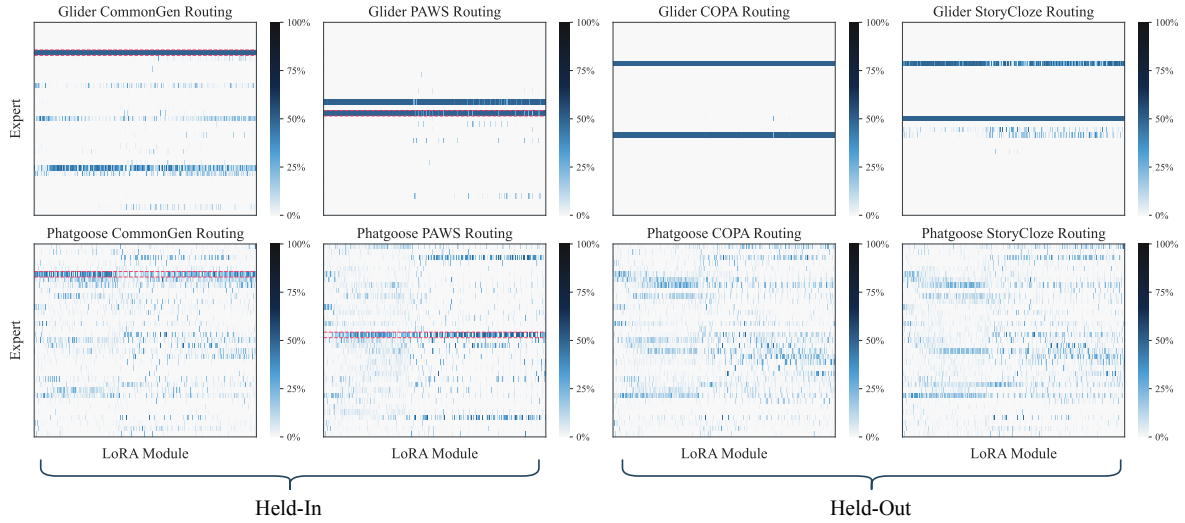

Figure 2: We present routing heatmaps for GLIDER and Phatgoose on two held-in and two held-out tasks. For held-in tasks, oracle experts are marked with red dashed lines. GLIDER selects oracle experts more frequently than Phatgoose for held-in tasks, leading to improvements of 3.3% on CommonGen and 6.5% on PAWS. For held-out tasks, GLIDER also tends to select the most relevant experts across most LoRA modules, resulting in improvements of 2.2% on COPA and 5.8% on StoryCloze.

(PEFT) (Liu et al., 2022; Sung et al., 2022; Poth et al., 2023) methods like LoRA (Hu et al., 2022) to train the expert models. Since PEFT typically has lower computational and communication costs than full-model finetuning (Hu et al., 2022; Liu et al., 2022), the use of PEFT makes it easier to participate and contribute. PEFT methods introduce modules throughout the model – for example, LoRA (Hu et al., 2022) introduces a low-rank update at every linear layer in the model. We refer to each of these updates as a *module*. Subsequently, the trained expert models and additional information are shared with the aggregators. The aggregator's job is to collect these expert models and the additional information and design the post-hoc routing mechanism. This mechanism will effectively direct incoming queries to the most appropriate expert model for each token and at each module to ensure optimal performance on both seen and unseen tasks. This approach allows for the seamless integration of new capabilities by adding expert models to the existing pool. Next, we formally define our contributor-aggregator framework.

Let us assume that there are N contributors, $\{c_1, c_2, \ldots, c_N\}$, and each contributor c_i has access to a task-specific datasets $\mathcal{D}_i$. Each contributor, c_i, follows the predefined training protocol $\mathcal{T}$ provided by the aggregator. The training protocol ($\mathcal{T}$) takes in a base model ($\boldsymbol{\theta}_{\text{base}}$) and a dataset ($\mathcal{D}_i$). It returns the expert model parameters (ϕ_i) along with any additional information (Ψ_i) that needs to be shared with the aggregators, for example, the gate vectors described in Section 4.1. Specifically, $\{\phi_i,\ \Psi_i\} \leftarrow \mathcal{T}(\boldsymbol{\theta}_{\text{base}}, \mathcal{D}_i)$. All contributors share this information with the aggregator, which creates a pool of models containing $\{(\phi_i, \Psi_i)\}_{i=1}^{N}$. The aggregators ($\mathcal{A}$) then uses these expert models and the auxiliary information to create a routing mechanism $\mathcal{R}(.)$ that takes the user query q as the input and return routing path describing how the information is routed through the given set of expert models. Formally, $\mathcal{R}(.) \leftarrow \mathcal{A}(\{(\phi_i, \Psi_i)\}_{i=1}^{N})$. The function $\mathcal{R}(.)$ describe the full path of input query by making various choices about 1) expert input granularity, choosing to route per-token, per-query, or per-task, 2) expert depth granularity, opting for either per-module or model-level routing, and 3) selecting between sparse or dense routing. Finally, the aggregator uses the routing mechanism to answer incoming queries.

4 Methodology

To recap, our goal is to build a MoErging method that dynamically routes queries to a diverse pool of specialized expert models, addressing the challenge of effectively handling queries from various tasks and ensuring both held-in and held-out performance. Our proposed method, **G**lobal and **L**ocal **I**nstruction **D**riven **E**xpert **R**outer (GLIDER), leverages a combination of local and global routing vectors to achieve this goal. Specifically, contributors train task-specific routing vectors, while an LLM generates global semantic task instructions, which are then converted to global instruction routing vectors. During inference, these local and global routing vectors are combined to perform top-k discrete routing, directing queries to the most suitable expert model. This process is visualized in Figure 1

and described in detail below.

4.1 Expert Training Protocol

Our expert training protocol $\mathcal{T}$ takes as input the base model parameters, θ_{base}, and a dataset $\mathtt{d}$ and performs three steps to obtain the required output. First, we train the LoRA experts (ϕ) and then the local routing vectors ($\mathtt{l}$) while keeping the LoRA experts fixed. Finally, we train the global routing vector ($\mathtt{g}$) by using an LLM and an embedding model. Formally, in our case, $\phi,\ \Psi = \{\mathtt{l}, \mathtt{g}\} \leftarrow \mathcal{T}(\theta_{\text{base}}, \mathtt{d})$ which are then shared with the aggregators to create the routing mechanism. We described these steps in detail below.

PEFT Training of Expert Model. GLIDER is compatible with expert models trained using parameter-efficient finetuning methods (*e.g.* LoRA (Hu et al., 2022), Adapters (Houlsby et al., 2019)) that introduce small trainable modules throughout the model. We focus on PEFT experts because they typically have lower computational and communication costs than full-model finetuning (Yadav et al., 2023a), making it easier to train and share expert models. Following Phatgoose (Muqeeth et al., 2024), this work specifically focuses on LoRA (Hu et al., 2022) due to its widespread use. LoRA introduces a *module* comprising the trainable matrices $\mathtt{B} \in \mathbb{R}^{\mathtt{d}\times\mathtt{r}}$ and $\mathtt{A} \in \mathbb{R}^{\mathtt{r}\times\mathtt{n}}$ in parallel to each linear layer with parameters $\mathtt{W} \in \mathbb{R}^{\mathtt{d}\times\mathtt{n}}$. Given the $\mathtt{i}^{\text{th}}$ input token activation $\mathtt{u_i}$, LoRA modifies the output of the linear layer from $\mathtt{Wu_i}$ to $\mathtt{Wu_i} + \frac{\alpha}{\mathtt{r}} \cdot \mathtt{BAu_i}$ where α is a constant and usually is set to 1. During training, the matrices $\mathtt{A}$ and $\mathtt{B}$ are trainable, while the original linear layer $\mathtt{W}$ is kept frozen. We denote the final trained expert parameters with $\phi = \{(\mathtt{A_1}, \mathtt{B_1}), \ldots, (\mathtt{A_m}, \mathtt{B_m})\}$, where $\mathtt{m}$ is the number of modules in the model.

Training Local Routing Vectors. Following Phatgoose (Muqeeth et al., 2024), after training the PEFT modules on their dataset, a local router is introduced before each PEFT module. This router, employing a shared vector across all queries and tokens, dynamically determines the utilization of the PEFT module based on the input token activations. The router is trained for a small number of steps using the same dataset and objective as the PEFT module while keeping the expert PEFT parameters fixed. This process effectively learns to associate the token activation patterns with the learned expert model. For LoRA, the local router, represented by a trainable vector $\mathtt{v} \in \mathbb{R}^{\mathtt{d}}$, controls the contribution of the PEFT module to the final output. This results in a modified linear layer of the form $\mathtt{Wu_i} + \frac{\alpha}{\mathtt{r}} \cdot \mathtt{BAu_i} \cdot \mathtt{sigmoid}(\mathtt{v}^\mathsf{T}\mathtt{u_i})$, where α, $\mathtt{W}$, $\mathtt{B}$, and $\mathtt{A}$ are frozen, and the local router $\mathtt{v}$ is learned. We denote the final local routing vectors as $\mathtt{l} = \{\mathtt{v_1}, \ldots, \mathtt{v_m}\}$ where $\mathtt{m}$ is the number of modules in the model.

Creating LLM-Aided Global Routing Vector. The local routing vectors capture the intricate relationships between token activations and expert models, enabling efficient query routing in cases where no dedicated expert is available. Conversely, for queries corresponding to held-in tasks, direct retrieval of the relevant expert model is preferred to process the full query. For this purpose, we create a global routing vector that utilizes an LLM to generate a semantically-informed instruction, termed as task description, which effectively captures the essence of the kind of queries the expert can handle. We prompt an LLM with three randomly selected in-context examples to generate this task description. We used the `gpt-4-turbo` model along with the prompt provided in Appendix B. The resulting task description is then embedded using an off-the-shelf embedding model, specifically the `nomic-embed-text-v1.5` model, to produce a global routing vector for the task. We denote the global routing vector as $\mathtt{g} \in \mathbb{R}^{\mathtt{d_g}}$.

4.2 GLIDER: Inference Expert Aggregation

Following training, all contributors share their expert models along with the auxiliary information comprising of the local and global routing vectors, $\{\phi^{\mathtt{t}},\ \mathtt{l}^{\mathtt{t}},\ \mathtt{g}^{\mathtt{t}}\}_{\mathtt{t}=1}^{\mathtt{N}}$, where $\mathtt{t}$ indexes the input tokens with the aggregators. The GLIDER method subsequently leverages this information to perform inference on arbitrary queries.

Local Router. Before each input module $\mathtt{m}$, a separate local router weight $\mathtt{L_m} \in \mathbb{R}^{\mathtt{N}\times\mathtt{d}}$ is inserted to make local per-token, per-module routing decisions. For a given module $\mathtt{m}$ and expert model $\mathtt{c}$, we have $\bar{\mathtt{v}}_{\mathtt{m}}^{\mathtt{c}} = \frac{\mathtt{v}_{\mathtt{m}}^{\mathtt{c}} - \mu(\mathtt{v}_{\mathtt{m}}^{\mathtt{c}})}{\sigma(\mathtt{v}_{\mathtt{m}}^{\mathtt{c}})}$, where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation respectively. Next, we obtain the local router for module $\mathtt{m}$ by stacking these standardised local routing vectors as $\mathtt{L_m} = [\bar{\mathtt{v}}_{\mathtt{m}}^{\mathtt{1}}; \ldots; \bar{\mathtt{v}}_{\mathtt{m}}^{\mathtt{N}}] \in \mathbb{R}^{\mathtt{N}\times\mathtt{d}}$. Next, for each token $\mathtt{i}$ with activation $\mathtt{u_i}$ coming into module $\mathtt{m}$, we standardise it to obtain $\bar{\mathtt{u}}_{\mathtt{i}} = \frac{\mathtt{u_i} - \mu(\mathtt{u_i})}{\sigma(\mathtt{u_i})}$. We then compute the local affinity scores, $\mathtt{s}_{\mathtt{m}}^{\mathtt{loc}} \in \mathbb{R}^{\mathtt{N}}$ between the local

router $\mathtt{L_m}$ and $\mathtt{u_i}$ as $\mathtt{s_m^{loc}} = \mathtt{cos\text{-}sim}(\mathtt{L_m}, \mathtt{u_i})$.

Global Router. The global router aims to capture task semantics to retrieve relevant experts for any given input query. We create the global router weight $\mathtt{G} \in \mathbb{R}^{\mathtt{N} \times \mathtt{d_g}}$ by stacking the global routing vectors from all the expert models as $\mathtt{G} = [\mathtt{g}^1; \ldots; \mathtt{g}^{\mathtt{N}}]$. This router is not a part of the base model and is added before the model to independently process the full query. Given an input query $\mathtt{u}$ along with three few-shot input-output pairs of similar queries, we prompt an LLM (`gpt-4-turbo`) using the template provided in Appendix B to obtain a task description for the query. We then embed this task description using the same embedding model (`nomic-embed-text-v1.5`) to obtain the vector $\mathtt{q_u} \in \mathbb{R}^{\mathtt{d_g}}$. We then compute the global affinity score, $\mathtt{s^{glob}} \in \mathbb{R}^{\mathtt{N}}$, by computing the cosine similarity as $\mathtt{s^{glob}} = \mathtt{cos\text{-}sim}(\mathtt{G}, \mathtt{q_u})$.

Combining Global and Local Router. At each module $\mathtt{m}$, we have the global and local affinity scores $\mathtt{s^{glob}}$ and $\mathtt{s_m^{loc}}$ respectively. Following Phatgoose (Muqeeth et al., 2024), we scale the local scores with a factor of $1/\sqrt{\mathtt{N}}$. However, the global router's main goal is to retrieve the correct expert for the held-in tasks. Therefore, we first check if the expert with the highest global affinity score ($\mathtt{max}(\mathtt{s^{glob}})$) is above a threshold ($\mathtt{p}$). If such experts exist, then we set a high α to enforce retrieval and vice versa. Hence, we propose to scale the global scores with α, where $\alpha = \gamma \cdot \mathbb{I}_{\{\mathtt{max}(\mathtt{s^{glob}}) - \mathtt{p} > 0\}} + \beta$, where $\mathtt{p}$ is the cosine similarity threshold, and γ and β are scaling hyperparameters. Using our ablation experiments in Section 5.4, we set $\mathtt{p} = 0.8$, $\gamma = 100$ and $\beta = 3$. We then obtain the final affinity score $\mathtt{s} \in \mathbb{R}^{\mathtt{N}} = \alpha \cdot \mathtt{s^{glob}} + \mathtt{s_m^{loc}}/\sqrt{\mathtt{N}}$. Then `GLIDER` selects the top-$\mathtt{k}$ experts after performing `softmax` over the final affinity score $\mathtt{s}$ as $\mathcal{E}_{\mathtt{top}} = \mathtt{top\text{-}k}(\mathtt{softmax}(\mathtt{s}))$. Finally, the output of the module for token activation u_i is computed as $\mathtt{Wu_i} + \sum_{\mathtt{k} \in \mathcal{E}_{\mathtt{top}}} \mathtt{s_k} \cdot \mathtt{B_k A_k u_i}$.

5 Experiments

5.1 Setting

Dataset. Our experiments utilize the multitask prompted training setup (**T0-HI**) introduced by Sanh et al. (2021), which has become a standard benchmark for evaluating held-in performance as well as generalization to unseen tasks (Chung et al., 2022; Longpre et al., 2023; Jang et al., 2023; Zhou et al., 2022). Phatgoose (Muqeeth et al., 2024) shows how local routing can be used for generalization to unseen domain, hence, following them, we employ LM-adapted T5.1.1 XL (Lester et al., 2021) as our base model which is a 3B parameter variant of T5 (Raffel et al., 2020) further trained on the C4 dataset using a standard language modeling objective. For held-out evaluations, we follow Phatgoose (Muqeeth et al., 2024) and use three held-out benchmark collections. We use the T0 held-out (**T0-HO**) datasets used in Sanh et al. (2021) and the two subsets of BIG-bench (BIG-bench authors, 2023). Specifically, we use BIG-bench Hard (**BBH**) (Suzgun et al., 2022), consisting of 23 challenging datasets, and BIG-bench Lite (**BBL**) (BIG-bench authors, 2023), a lightweight 24-dataset proxy for the full benchmark. Similar to Muqeeth et al. (2024), we exclude certain BIG-bench datasets due to tokenization incompatibility with the T5 tokenizer.

Expert Creation. To create the pool of expert module for routing, we follow Muqeeth et al. (2024) and use two distinct dataset collections: ❶ T0 Held-In (Sanh et al., 2021) consisting of the 36 held-in prompted datasets for tasks from the T0 training procedure. ❷ The "FLAN Collection" (Longpre et al., 2023) which significantly expands the T0 tasks by incorporating prompted datasets from SuperGLUE (Wang et al., 2019a), Super Natural Instructions (Wang et al., 2022b), dialogue datasets, and Chain-of-Thought datasets (Wei et al., 2022b). Following Muqeeth et al. (2024), we create `166` specialized models from the FLAN Collection. For each dataset in these collections, we train Low-Rank Adapters (LoRAs) (Hu et al., 2021) modules resulting in pools of `36` and `166` expert models for T0 Held-In and FLAN, respectively. Similar to Phatgoose, we use a rank of $r = 16$ and train for `1000` steps using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-3} and a warmup ratio of 0.06. After training the LoRA module, we freeze it and train the local routing vectors for an additional 100 steps with the same hyperparameters. Finally, following prior work (Shazeer et al., 2016; Du et al., 2022; Lepikhin et al., 2020), `GLIDER` performs top-k routing with $k = 2$.

5.2 Baselines

Expert Merging. Model Merging (Yadav et al., 2023b; Choshen et al., 2022) involves averaging

Table 1: Performance evaluated on the T0 set and FLAN set. We present the performance on both held-in tasks (*i.e.* T0-HI) and held-out tasks (*i.e.* T0-HO, BBH, and BBL). We compare the following methods: (1) performance upper bound, *i.e.* Oracle Expert; (2) zero-shot baselines, *i.e.* Multi-Task Fine-Tuning, Expert Merging, Arrow, and Phatgoose; (3) few-shot baselines, *i.e.* LoRA Hub and GLIDER. We mark the best performance besides the upper bound (*i.e.*, Oracle Expert) in **bold**.

Method	T0				FLAN	
	T0-HI	T0-HO	BBH	BBL	BBH	BBL
Oracle Expert	69.60	51.60	34.90	36.60	38.90	45.40
Multi-Task Fine-Tuning	55.90	51.60	34.90	36.60	**38.90**	**45.40**
Expert Merging	30.73	45.40	**35.30**	36.00	34.60	34.00
Arrow	39.84	55.10	33.60	34.50	30.60	29.60
Phatgoose	61.42	56.90	34.90	37.30	35.60	35.20
LoRA Hub	31.90	46.85	31.35	31.18	34.50	30.54
GLIDER	**68.04**	**57.78**	35.29	**37.46**	35.07	35.52

the parameters of multiple models or modules to create a single aggregate model. We merge by multiplying the LoRA matrices and then taking an unweighted average of all the experts within the pool. It is important to note that this merging strategy requires homogeneous expert module architectures; in contrast, GLIDER can accommodate heterogeneous expert modules.

Arrow. Following Ostapenko et al. (2024), we employ a routing mechanism where gating vectors are derived from LoRA expert modules. Specifically, the first right singular vector of the outer product of each module's LoRA update (BA) serves as its gating vector. Input routing is determined by a probability distribution based on the absolute dot product between the input representation and each gating vector. We utilize top-k routing with $k = 2$.

Phatgoose. Phatgoose (Muqeeth et al., 2024) first learn the LoRA modules for each, followed by learning a sigmoid gating vector similar to our local router. During inference, they make routing decisions for each token independently for all modules. Specifically, they first standardize the input token activations and gating vectors from all experts and then perform similarity-based top-2 routing.

LoRA Hub. LoraHub (Huang et al., 2023) method performs gradient-free optimization using few-shot task samples to learn mixing coefficients for different expert models while keeping them fixed. Once the coefficients are learned, they merge the experts with the learned weight and route through the merged expert.

Multi-task Fine-Tuning. Multitask training is a proven method for enhancing zero-shot generalization (Sanh et al., 2021; Wei et al., 2022a) but is infeasible given our problem setting and data access

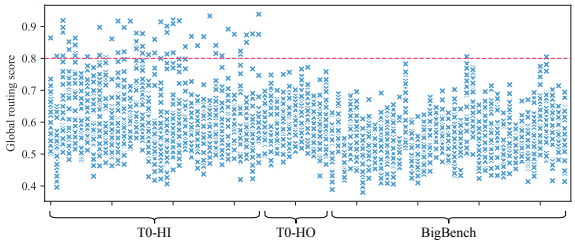

Figure 3: Global routing scores for tasks in the T0 set. The red horizontal line indicates our design threshold of 0.8. Each column represents an evaluated task from T0-HI, T0-HO, BigBench using T0 held-in experts. All global routing scores for each task are plotted, corresponding to the 35 experts in total.

limitations. We include it as a baseline using publicly available models. Specifically, we utilize the T0-3B model (Sanh et al., 2021) for the T0 Held-In datasets, given its training on a matching dataset collection. For FLAN, a directly comparable publicly available model is unavailable; therefore, we report FLAN-T5 XL results trained on a different, undisclosed dataset mixture, while acknowledging the limitations of this indirect comparison.

Oracle. Following (Jang et al., 2023) and (Muqeeth et al., 2024), we employ an Oracle routing scheme as a performance upper bound. This scheme selects the expert exhibiting optimal performance on a given evaluation dataset, thus representing a non-zero-shot approach.

5.3 Main Results

Table 1 presents the comparison results among our GLIDER and six baselines on both held-in and held-out settings. We report the average performance across all tasks for each setting, please see Appendix D for each tasks metric. To further illustrate the performance, we also include the results of Oracle Expert, which has extra access to the task identities of expert modules and evaluated datasets and can be regarded as an *upper bound*.

T0 Setting. In the T0 task set, the following observations can be drawn: ❶ For the held-in tasks, *i.e.* T0-HI, GLIDER significantly outperforms other baselines and almost matches the performance of Oracle Expert upper bound. ❷ For T0-HO and BBL tasks, GLIDER achieves the best performance among all the methods, including Oracle Expert upper bound. ❸ GLIDER has negligible lower performance, *i.e.* 0.01%, compared to the Expert Merging baseline in BBH but outperforms it by around 12% on T0-HO and 1.5% on BBL. Besides Expert Merging, GLIDER outperforms all other methods on BBH, including the Oracle Expert upper bound.

Table 2: Ablation on the instruction coefficient α. We mark the best performance in **bold** and the performance corresponding to the selected α by GLIDER in blue.

α	T0			
	T0-HI	**T0-HO**	**BBH**	**BBL**
0	61.42	56.90	34.90	37.30
1	62.20	57.04	35.05	**37.79**
3	63.40	57.78	**35.29**	37.46
10	65.52	**57.98**	34.80	37.04
100	**68.04**	53.22	31.73	34.97
1000	66.88	52.91	30.71	34.31
3000	66.69	52.37	30.03	33.24

The key insight is that held-in tasks benefit from explicit task identification (via global router) while held-out tasks require compositional reasoning (via local router). Our multi-scale approach allows both modes to coexist.

5.4 Ablation Study and Further Investigation

Ablation on the global routing scale α. To illustrate how the specialization and generalization abilities change as we scale the coefficient α of the global routing score, we conduct the ablation study of α ranging $\{1, 3, 10, 100, 1000, 3000\}$. As shown in Table 2, we present experimental results of the T0 task set on both held-in and held-out tasks. For held-in tasks, *i.e.* T0-HI, GLIDER can select the optimal α to scale the global routing score. For held-out tasks, *i.e.* {T0-HO, BBH, BBL}, GLIDER produce either the optimal α (for BBH) or the sub-optimal α with slightly lower performance to the optimal ones (for T0-HO and BBL). Lastly, note that Phatgoose correspond to the setting where there is no global semantics used, i.e., $\alpha = 0$.

Ablation on the routing strategy. There exists a trade-off between performance and efficiency when using different `top-k` routing strategies (Ramachandran and Le, 2019). To investigate the impact of routing strategy in GLIDER, we evaluate `top-k` routing of `k` in $\{1, 2, 3\}$. Moreover, we further evaluate the `top-p` routing (Huang et al., 2024c; Zeng et al., 2024) of `p` in $\{25\%, 50\%, 75\%\}$, where each token selects experts with higher routing probabilities until the cumulative probability exceeds threshold `p`. As shown in Table 3, we can draw the following conclusions: (1) For `top-k` routing, `k` $= 2$ shows comparable or better performance than `k` $= 3$, particularly for T0-HO and BBH, while offering improved efficiency. (2) For `top-p` routing, higher `p` values consistently yield better performance at the cost of efficiency. Therefore, we use `top-2` routing in GLIDER by default.

Table 3: Ablation on the routing strategy. GLIDER employs `top-2` routing. We mark the best performance among `top-k` and `top-p` routing in **bold**, respectively.

Method	T0			
	T0-HI	**T0-HO**	**BBH**	**BBL**
Top-1	67.96	56.07	33.91	35.82
Top-2	68.04	**57.78**	**35.39**	37.46
Top-3	**68.06**	57.52	35.08	**38.55**
Top-25%	67.98	56.53	34.10	36.32
Top-50%	67.95	57.25	35.07	37.49
Top-75%	**68.02**	**57.86**	**35.38**	**38.65**

Investigation on the threshold design of global scores. As in Section 4, we compute the scale α for global scores using the formula $\alpha = \gamma * \mathbb{I}_{\{\max(s^{glob})-0.8>0\}} + \beta$, where we establish a threshold of 0.8 to differentiate evaluated tasks. Figure 3 presents the global routing scores for each task in the T0 set to motivate the rationale behind this design. For all held-in tasks (*i.e.*, T0-HI), at least one expert (typically the oracle expert trained on the evaluated task) achieves global routing scores exceeding 0.8. Consequently, GLIDER applies a higher $\alpha = 100$, enabling effective identification of tasks corresponding to a specifically trained expert and enhancing retrieval of this oracle expert. For nearly all held-out tasks (*i.e.*, T0-HO and BigBench), no global routing score surpasses 0.8, prompting GLIDER to utilize a lower $\alpha = 3$. Two exceptions among the held-out tasks are `bbq_lite_json` and `strange_stories` in BigBench, where one score marginally exceeds 0.8 in each case. For these two, GLIDER employs the higher $\alpha = 100$, resulting in performance improvements of 1.3% and 2.9% respectively over $\alpha = 3$, thus showing the effectiveness of our design.

6 Conclusion

This paper introduces GLIDER, a novel multi-scale routing mechanism that incorporates both global semantic and local token-level routers. By leveraging the semantic reasoning capabilities of LLMs for global expert selection and refining these choices with a learned local router, GLIDER addresses the limitations of existing methods that often perform poorly on held-in tasks. Our empirical evaluation on T0 and FLAN benchmarks, using T5-based experts, demonstrates that GLIDER achieves substantial improvements in held-in task performance while maintaining competitive generalization on held-out tasks. These findings suggest that incorporating global semantic task context into routing mechanisms is crucial for building robust and practically useful routing-based systems.

7 Limitation

The main limitation of GLIDER lies in its heavy dependence on large language models (specifically GPT-4) for generating semantic task descriptions. This reliance introduces potential accessibility barriers due to API costs. Furthermore, investigating the application of GLIDER to other modalities beyond language tasks, such as vision or multi-modal expert models, could unlock new capabilities for specialized model routing.

References

Wikiquote, russian proverbs.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. *arXiv preprint*.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439.

Joshua Ackerman and George Cybenko. 2020. A survey of neural networks and formal languages. *arXiv preprint*.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual question answering. *arXiv preprint*.

Akshay Agrawal, Shane Barratt, and Stephen Boyd. 2020. Learning convex optimization models. *arXiv preprint*.

Scott Alexander. 2020. A very unlikely chess game.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA. Association for Computing Machinery.

Miltiadis Allamanis, Earl T. Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *ACM Comput. Surv.*, 51(4).

Miltiadis Allamanis, Pankajan Chanthirasegaran, Pushmeet Kohli, and Charles Sutton. 2016. Learning continuous semantic representations of symbolic expressions. *arXiv preprint*.

Uri Alon, Shaked Brody, Omer Levy, and Eran Yahav. 2018. code2seq: Generating sequences from structured representations of code. *arXiv preprint*.

Uri Alon, Roy Sadaka, Omer Levy, and Eran Yahav. 2020. Structural language models of code. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 245–256. PMLR.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Prithviraj Ammanabrolu, William Broniec, Alex Mueller, Jeremy Paul, and Mark O. Riedl. 2019. Toward automated quest generation in text-adventure games. *arXiv preprint*.

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. *arXiv preprint*.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint*.

Brandon Amos and J. Zico Kolter. 2017. Optnet: Differentiable optimization as a layer in neural networks. *arXiv preprint*.

Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using BERT sentence embedding for humor detection. *arXiv preprint*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact

checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Salvatore Attardo. 2017. Humor in language. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*.

R H. Baayen, R Piepenbrock, and L Gulikers. 1995. Celex2 ldc96l14.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? Learning to discriminate machine from human generated text. *arXiv preprint*.

Matej Balog, Alexander L. Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. 2016. Deepcoder: Learning to write programs. *arXiv preprint*.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI'03: Proceedings of the 18th International Joint Conference on Artificial Intelligence*, page 805–810, San Francisco. Morgan Kaufmann.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Oren Barkan and Noam Koenigstein. 2016. ITEM2VEC: Neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Piscataway, NJ. Institute of Electrical and Electronics Engineers.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Sumit Basu and Janara Christensen. 2013. Teaching classification boundaries to humans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 109–115, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. *arXiv preprint*.

Nihat Bayat and Gökhan Çetinkaya. 2020. The relationship between inference skills and reading comprehension. *TED EĞİTİM VE BİLİM (Education and Science)*, 45(203):177–190.

Mayur J. Bency, Ahmed H. Qureshi, and Michael C. Yip. 2019. Neural path planning: Fixed time, near-optimal path generation via oracle imitation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3965–3972, Piscataway, NJ. Institute of Electrical and Electronics Engineers.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Jean Berko. 1958. The child's learning of english morphology. WORD, 14(2-3):150–177.

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint*.

Gregor Betz, Christian Voigt, and Kyle Richardson. 2020. Critical thinking for language models. *arXiv preprint*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint*.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? Try ARC-DA, the direct-answer AI2 reasoning challenge. *arXiv preprint*.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020a. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, Online. Association for Computational Linguistics.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020b. On the practical ability of recurrent neural networks to recognize hierarchical languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1481–1494, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, and Pushmeet Kohli. 2017. Deep API programmer: Learning to program with APIs. *arXiv preprint*.

Alan W. Biermann. 1978. The inference of regular LISP programs from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(8):585–600.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *arXiv preprint*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7423–7439, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Large scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Yulia V. Bodrova. 2007. *Russian Proverbs and Sayings and Their English Equivalents*. AST, Moscow.

Nicholas Boillot. 2019. Vector forms as a foreign language.

Paul F. Boller, Jr. and John George. 1989. *They Never Said It: A Book of Fake Quotes, Misquotes, and Misleading Attributions*. Oxford University Press, Oxford.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint*.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Matko Bošnjak, Tim Rocktäschel, Jason Naradowsky, and Sebastian Riedel. 2017. Programming with a differentiable Forth interpreter. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 547–556.

Gwern Branwen. 2020. GPT-3 creative fiction. *Gwern.net*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

Tom B. 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 M. 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.

Thomas Bugnyar, Stephan A. Reber, and Cameron Buckner. 2016. Ravens attribute visual access to unseen competitors. *Nature Communications*, 7:article 10506.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

Corrado Böhm. 1964. On a family of Turing machines and the related programming language. *ICC Bulletin*, 3:187–194.

Lucas Caccia, Edoardo Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni. 2023. Multi-head adapter routing for cross-task generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kate Cain and Jane V. Oakhill. 1999. Inference making ability and its relation to comprehension failure. *Reading and Writing*, 11(5–6):489–503.

Maya Cakmak and Andrea L. Thomaz. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence*, 217:198–215.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Josep Call and Michael Tomasello. 2008. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12:187–192.

Tracy Canfield. 2010. Machine translation of Klingon.

Nathanael Chambers. 2012. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 98–106, Jeju Island, Korea. Association for Computational Linguistics.

Sharath Chandra Guntuku, Mingyang Li, Louis Tay, and Lyle H. Ungar. 2019. Studying cultural differences in emoji usage across the East and the West. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 226–235, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Nick Chater and Paul Vitányi. 2003. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7:19–22.

Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. 2020. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019a. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539, Piscataway, NJ. Institute of Electrical and Electronics Engineers.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Ricson Chen. 2020. Transformers play chess.

Xinyun Chen, Chang Liu, and Dawn Song. 2019b. Execution-guided neural program synthesis. https://openreview.net/pdf?id=H1gfOiAqYm.

Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. 2024. DAM: Dynamic adapter merging for continual video qa learning. *arXiv preprint arXiv:2403.08755*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint*.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *arXiv preprint*.

François Chollet. 2019. On the measure of intelligence.

François Chollet. 2020. Abstraction and reasoning challenge.

Noam Chomsky and Marcel P. Schützenberger. 1959. The algebraic theory of context-free languages. In P. Braffort and D. Hirschberg, editors, *Computer Programming and Formal Systems*, volume 26 of *Studies in Logic and the Foundations of Mathematics*, pages 118–161. Elsevier.

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*.

Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. 2023. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. 2017. CycleGAN, a master of steganography. *arXiv preprint*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Rudi L. Cilibrasi and Paul M.B. Vitanyi. 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint*.

Robert J. Clark and Robert R. Jackson. 1994. Self recognition in a jumping spider: Portia labiata females discriminate between their own draglines and those of conspecifics. *Ethology Ecology & Evolution*, 6(3):371–375.

Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, María José Ramírez-Quintana, and Susumu Katayama. 2020. Automated data transformation with inductive programming and dynamic background knowledge. In *Machine Learning and Knowledge Discovery in Databases*, pages 735–751, Cham. Springer.

Lidia Contreras-Ochando, César Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, María José Ramírez-Quintana, and Susumu Katayama. 2018. General-purpose declarative inductive programming with domain-specific background knowledge for data wrangling automation. *arXiv preprint*.

Irving M. Copi, Carl Cohen, and Victor Rodych. 2018. *Introduction to Logic*. Taylor & Francis.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint*.

Kate Crawford. 2017. The trouble with bias. https://www.youtube.com/watch?v=fMym_BKWQzk. Keynote address, NIPS 2017, Long Beach CA. Dec. 5, 2017.

Andrew Cropper, Rolf Morel, and Stephen Muggleton. 2020. Learning higher-order logic programs. *Machine Learning*, 109:1289–1322.

Andrew Cropper and Stephen H. Muggleton. 2016. Metagol system.

Andrew Cropper, Alireza Tamaddoni-Nezhad, and Stephen H. Muggleton. 2016. Meta-interpretive learning of data transformation programs. In *Inductive Logic Programming*, pages 46–59, Cham. Springer.

Joe Cruse. 2015. Emoji usage in TV conversation. *Twitter blog*.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A learning environment for text-based games. *arXiv preprint*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jim Daley. 2021. White Chicago cops use force more often than Black officers. *Scientific American*.

Sahith Dambekodi, Spencer Frazier, Prithviraj Ammanabrolu, and Mark O. Riedl. 2020. Playing text-based games with common sense. *arXiv preprint*.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2017. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 48–54, Valencia, Spain. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127:591–621.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, and Pushmeet Kohli. 2017. RobustFill: Neural program learning under noisy I/O. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 990–998, New York, NY, USA. Association for Computing Machinery.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv preprint*.

Kaustubh Dhole, Gurdeep Singh, Priyadarshini P. Pai, and Sukanta Mondal. 2014. Sequence-based prediction of protein–protein interaction sites with l1-logreg classifier. *Journal of Theoretical Biology*, 348:47–54.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and T. Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models' memories. In *Annual Meeting of the Association for Computational Linguistics*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Tiansi Dong, Chengjiang Li, Christian Bauckhage, Juanzi Li, Stefan Wrobel, and Armin B. Cremers. 2020. Learning syllogism with Euler neural-networks. *arXiv preprint*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Jon Durbin. 2024. airoboros: Customizable implementation of the self-instruct paper. `https://github.com/jondurbin/airoboros`.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint*.

Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. 2020. How can self-attention networks recognize Dyck-n languages? *arXiv preprint*.

Bora Edizel, Aleksandra Piktus, Piotr Bojanowski, Rui Ferreira, Edouard Grave, and Fabrizio Silvestri. 2019. Misspelling oblivious word embeddings. *arXiv preprint*.

Daniel Edmiston and Karl Stratos. 2018. Compositional morpheme embeddings with affixes as functions and stems as arguments. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 1–5, Melbourne, Australia. Association for Computational Linguistics.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liat Ein Dor, Alon Halfon, Yoav Kantor, Ran Levy, Yosi Mass, Ruty Rinott, Eyal Shnarch, and Noam Slonim. 2018. Semantic relatedness of Wikipedia concepts – benchmark data and a working solution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Ahmed El-Kishky, Frank Xu, Aston Zhang, and Jiawei Han. 2019. Parsimonious morpheme segmentation with an application to enriching word embeddings. *arXiv preprint*.

Ran El-Yaniv and David Yanay. 2013. Semantic sort: A supervised approach to personalized semantic relatedness.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *arXiv preprint*.

Kevin Ellis and Sumit Gulwani. 2017. Learning to learn programs from examples: Going beyond program structure. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1638–1645.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2020. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *arXiv preprint*.

Richard Evans, Jose Hernandez-Orallo, Johannes Welbl, Pushmeet Kohli, and Marek Sergot. 2019. Making sense of sensory input. *arXiv preprint*.

Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. 2018. Can neural networks understand logical entailment? *arXiv preprint*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2020. Text editing by command. *arXiv preprint*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Xiaochao Fan, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120).

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

John K. Feser, Swarat Chaudhuri, and Isil Dillig. 2015. Synthesizing data structure transformations from input-output examples. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '15*, page 229–239, New York, NY, USA. Association for Computing Machinery.

Susan T. Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48:621–628.

Dawn P. Flanagan and Shauna G. Dixon. 2014. The Cattell-Horn-Carroll theory of cognitive abilities. In *Encyclopedia of Special Education*. John Wiley & Sons, Ltd.

Pierre Flener and Ute Schmid. 2008. An introduction to inductive programming. *Artificial Intelligence Review*, 29:45–62.

Jerry A. Fodor. 1975. *The Language of Thought*. Harvard University Press, Cambridge, MA.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Mark Forsyth. 2014. *The Elements of Eloquence: Secrets of the Perfect Turn of Phrase*. Berkley, New York.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR.

Lea Frermann, Shay B. Cohen, and Mirella Lapata. 2018. Whodunnit? Crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6:1–15.

Kathryn J. Friedlander and Philip A. Fine. 2018. "The penny drops": Investigating insight through the medium of cryptic crosswords. *Frontiers in Psychology*, 9.

Martins Frolovs. 2019. Teaching GPT-2 transformer a sense of humor: How to fine-tune large transformer models on a single GPU in PyTorch. *Towards Data Science, Medium*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure: A meta evaluation of factuality in summarization. *arXiv preprint*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical Intelligence*, page 1606–1611, San Francisco. Morgan Kaufmann.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint*.

Artur d'Avila Garcez and Luis C. Lamb. 2020. Neurosymbolic AI: The 3rd wave. *arXiv preprint*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer

Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint*.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 174–182, Athens, Greece. Association for Computational Linguistics.

Alexander L. Gaunt, Marc Brockschmidt, Rishabh Singh, Nate Kushman, Pushmeet Kohli, Jonathan Taylor, and Daniel Tarlow. 2016. TerpreT: A probabilistic programming language for program induction. *arXiv preprint*.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference, WWW '19*, page 514–525, New York, NY, USA. Association for Computing Machinery.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Dedre Gentner, Mary Jo Rattermann, and Kenneth D. Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25(4):524–575.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in English dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323. Special issue: Third International Conference on Computing and Network Communications (CoCoNet'19).

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020a. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020b. Transformer feed-forward layers are key-value memories. *arXiv preprint*.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. 2020. Irony detection in a multilingual context. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science*, volume 12036. Springer, Cham.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. *arXiv preprint*.

Sayan Ghosh and Shashank Srivastava. 2021. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. *CoRR*, arXiv:2109.06838.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.

Matthew L. Ginsberg. 2014. Dr.Fill: Crosswords and an implemented solver for singly weighted CSPs. *CoRR*, arXiv:1401.4597.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint*.

Arthur S. Goldberger. 1972. Structural equation methods in the social sciences. *Econometrica*, 40(6):979–1001.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20:818–829.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tür. 2020. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? An empirical study. In *Proc. Interspeech 2020*, pages 911–915.

Andrew S. Gordon. 2010. Choice of plausible alternatives (COPA).

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing machines. *arXiv preprint*.

Alex Graves, Greg Wayne, Malcom Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471—476.

C. Cordell Green, Richard J. Waldinger, David R. Barstow, Robert Elschlager, Douglas B. Lenat, Brian P. McCune, David E. Shaw, and Louis I. Steinberg. 1974. Progress report on program-understanding systems (AIM-240).

Cordell Green. 1981. Application of theorem proving to problem solving. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 202–222. Morgan Kaufmann.

H. Paul Grice and Peter F. Strawson. 1956. In defense of a dogma. *The Philosophical Review*, 65(2):141–158.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. 2019. Stochastic optimization of sorting networks via continuous relaxations. *arXiv preprint*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Sumit Gulwani. 2011. Automating string processing in spreadsheets using input-output examples. *SIGPLAN Not.*, 46(1):317–330.

Sumit Gulwani, William R. Harris, and Rishabh Singh. 2012. Spreadsheet data manipulation using examples. *Commun. ACM*, 55(8):97–105.

Sumit Gulwani, José Hernández-Orallo, Emanuel Kitzelmann, Stephen H. Muggleton, Ute Schmid, and Benjamin Zorn. 2015. Inductive programming meets the real world. *Commun. ACM*, 58(11):90–99.

Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. 2017a. Program synthesis. *Foundations and Trends in Programming Languages*, 4(1–2):1–119.

Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. 2017b. *Program Synthesis*. NOW, Boston.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-QA: A benchmark dataset for understanding disfluencies in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3309–3319, Online. Association for Computational Linguistics.

Shashank Gupta, Subhabrata Mukherjee, Krishan Subudhi, Eduardo Gonzalez, Damien Jose, Ahmed H Awadallah, and Jianfeng Gao. 2022. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arXiv:2204.07689*.

Isidor S. Gvarjalaże and Dzhuansher I. Mchedlishvili. 1971. *English Proverbs and Sayings*. Vysshaya shkola, Moscow.

Samuel Gyasi Obeng. 1996. The proverb as a mitigating and politeness strategy in Akan discourse. *Anthropological Linguistics*, 38(3):521–549.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Joseph Y. Halpern. 2016. *Actual causality*. MIT Press, Cambridge, MA.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020. ECONET: Effective continual pretraining of language models for event temporal reasoning. *arXiv preprint*.

Maria Hanzén. 2007. When in Rome, do as the Romans do: Proverbs as a part of EFL teaching. Master's thesis, Jönköping University, School of Education and Communication, Jönköping.

Yiding Hao, William Merrill, Dana Angluin, Robert Frank, Noah Amsel, Andrew Benz, and Simon Mendelsohn. 2018. Context-free transductions with neural stacks. *arXiv preprint*.

Francesca G.E. Happé. 1994. An advanced test of theory of mind: Understanding of story characters thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24:129–154.

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Irene Heim. 1983. On the projection problem for presuppositions. In Paul Portner and Barbara H. Partee, editors, *Formal Semantics - The Essential Readings*, pages 249–260. Blackwell, Oxford.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *arXiv preprint*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, Cham. Springer.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021a. Measuring coding challenge competence with APPS. *arXiv*, arXiv:2105.09938.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *arXiv preprint*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint*.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint*.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015a. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015b. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint*.

Álvaro Hernandez, Suhan Woo, Héctor Corrales, Ignacio Parra, Euntai Kim, D. Fernández Llorca, and Miguel A. Sotelo. 2020. 3D-DEEP: 3-dimensional deep-learning based on elevation patterns for road scene interpretation. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, Piscataway, NJ. Institute of Electrical and Electronics Engineers.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. 2020. RNNs can generate bounded hierarchical languages with optimal memory. *arXiv preprint*.

Mireille Hildebrandt. 2018. Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170355.

Keith J. Holyoak. 2012. Analogy and relational reasoning. In Keith J. Holyoak and Robert G. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, Oxford.

Richard P. Honeck. 1997. *A Proverb in Mind: The Cognitive Science of Proverbial Wit and Wisdom*. Lawrence Erlbaum Associates, Mahwah, NJ.

Alexandra Horowitz. 2017. Smelling themselves: Dogs investigate their own odours longer when modified in an "olfactory mirror" test. *Behavioural Processes*, 143:17–24.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, Georgia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799.

China Household Management Research Center, Ministry of Public Security. 2019. National name report 2018. http://news.cpd.com.cn/n18151/201901/t20190130_830962.html (Accessed 3 March 2021).

China Household Management Research Center, Ministry of Public Security. 2020. National name report 2019. https://www.mps.gov.cn/n2254314/n6409334/c6874817/content.html (Accessed 3 March 2021).

China Household Management Research Center, Ministry of Public Security. 2021. National name report 2020. https://www.mps.gov.cn/n2253534/n2253535/c7725981/content.html (Accessed 3 March 2021).

Hrisztalina Hrisztova-Gotthardt and Melita Aleksa Varga. 2015. *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*. De Gruyter Open, Warsaw.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024a. Lorahub: Efficient cross-task generalization via dynamic lora composition. *Preprint*, arXiv:2307.13269.

Daniel Huang, Prafulla Dhariwal, Dawn Song, and Ilya Sutskever. 2018. Gamepad: A learning environment for theorem proving. *arXiv preprint*.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. 2024b. Copa: General robotic manipulation through spatial constraints of parts with foundation models. *Preprint*, arXiv:2403.08248.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Songfang Huang, and Yansong Feng. 2024c. Harder tasks need more experts: Dynamic routing in moe models. *Preprint*, arXiv:2403.07652.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *arXiv preprint*.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 581–589, Prague, Czech Republic. Association for Computational Linguistics.

David Hume. 1739–1740. *A Treatise of Human Nature*. John Noon, London.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Annamarie W. Huttunen, Geoffrey K. Adams, and Michael L. Platt. 2017. Can self-awareness be taught? Monkeys pass the mirror test – again. *Proceedings of the National Academy of Sciences*, 114(13):3281–3283.

David Huynh and Stefano Mazzocchi. 2012. OpenRefine. `https://openrefine.org/`.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Instagram Engineering. 2015. Emojineering part 1: Machine learning for emoji trends. *Medium*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint*.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint*.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. *arXiv preprint arXiv:2302.03202*.

Mario Jarmasz. 2012. Roget's Thesaurus as a lexical resource for natural language processing. Master's thesis, University of Ottawa, Ottawa.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602, Online. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. *arXiv:1707.06209v1*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint*.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. *arXiv preprint*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, arXiv:1705.03551.

Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15*, volume 1, page 190–198, Cambridge, MA, USA. MIT Press.

Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6505–6520, Online. Association for Computational Linguistics.

Yuji Kanagawa and Tomoyuki Kaneko. 2019. Rogue-Gym: A new challenge for generalization in reinforcement learning. In *2019 IEEE Conference on Games (CoG)*, pages 1–8, Piscataway, NJ. Institute of Electrical and Electronics Engineers.

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, page 3363–3372, New York, NY, USA. Association for Computing Machinery.

Immanuel Kant. 1781/1787. *Critique of Pure Reason*. The Cambridge Edition of the Works of Immanuel Kant, edited by Paul Guyer and Allen W. Wood. Cambridge University Press.

Immanuel Kant. 1783. *Prolegomena to Any Future Metaphysics*, 2nd edition. Cambridge Texts in the History of Philosophy, edited by Gary Hatfield. Cambridge University Press.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andrej Karpathy. 2015. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy's blog*.

Lauri Karttunen. 2012. Simple and phrasal implicatives. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 124–131, Montréal, Canada. Association for Computational Linguistics.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? *arXiv preprint*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint*.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents.

Faisal Khan, Bilge Mutlu, and Jerry Zhu. 2011. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2020a. ParsiNLU: A suite of language understanding challenges for persian. *arXiv preprint*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020b. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.

Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. 2019. Cooperation and codenames: Understanding natural language processing via codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 160–166, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2015. Character-aware neural language models. *arXiv preprint*.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

Christo Kirov and Ryan Cotterell. 2018. Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Emanuel Kitzelmann. 2010. Inductive programming: A survey of program synthesis techniques. In *Approaches and Applications of Inductive Programming*, pages 50–73, Berlin. Springer.

Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis*, 63.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Jan Kocoń, Piotr Miłkowski, and Kamil Kanclerz. 2021. MultiEmo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. In *Computational Science – ICCS 2021*, pages 297–312, Cham. Springer.

Alexander W. Kocurek and Ethan Jerzak. 2021. Counterlogicals as counterconventionals. *Journal of Philosophical Logic*, 50:673–704.

Alexander W. Kocurek, Ethan Jerzak, and Rachel Etta Rudolph. 2020. Against conventional wisdom. *Philosophers' Imprint*, 20(22):1–27.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 62, New York, NY, USA. Association for Computing Machinery.

Jarmo Korhonen. 2009. Sprichwörter und zweisprachige lexikographie: Deutsch-schwedische und deutsch-finnische wörtebücher im vergleich. In C. Földes, editor, *Phraseologie disziplinär und interdisziplinär*, pages 537–549. Gunter Narr Verlag.

Dimitrios Kotsakos, Theodoros Lappas, Dimitrios Kotzias, Dimitrios Gunopulos, Nattiya Kanhabua, and Kjetil Nørvåg. 2014. A burstiness-aware approach for document dating. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14*, page 1003–1006, New York, NY, USA. Association for Computing Machinery.

Samuel Kounev, Jeffrey O. Kephart, Aleksandar Milenkoski, and Xiaoyun Zhu, editors. 2017. *Self-Aware Computing Systems*. Springer, Cham.

Sarah E. Kreps, Miles McCain, and Miles Brundage. 2020. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *SSRN*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *arXiv preprint*.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Niklas Kühl, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. 2020. Human vs. supervised machine learning: Who learns patterns faster? *arXiv preprint*.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. 2020. The NetHack learning environment. *arXiv preprint*.

Kevin Lacker. 2020. Giving GPT-3 a Turing test. *Kevin Lacker's blog*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint*.

Brenden M. Lake and Gregory L. Murphy. 2020. Word meaning in minds and machines. *arXiv preprint*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

George Lakoff and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago Press, Chicago.

Yair Lakretz, Théo Desbordes, Jean-Rémi King, Benoît Crabbé, Maxime Oquab, and Stanislas Dehaene. 2021a. Can RNNs learn recursive nested subject-verb agreements? *arXiv preprint*.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021b. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Lample and François Charton. 2019. Deep learning for symbolic mathematics. *arXiv preprint*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Remi Lebret, David Grangier, and Michael Auli. 2016a. Neural text generation from structured data with application to the biography domain. *Preprint*, arXiv:1603.07771.

Rémi Lebret, David Grangier, and Michael Auli. 2016b. The wikibio corpus: A corpus of biographical texts for natural language generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: A research direction. *arXiv preprint*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

Iddo Lev, Bill MacCartney, Christopher Manning, and Roger Levy. 2004. Solving logic puzzles: From robust processing to precise semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Ran Levy, Liat Ein-Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. 2015. TR9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 419–424, Beijing, China. Association for Computational Linguistics.

Sharon Levy, Michael Saxon, and William Yang Wang. 2021. Investigating memorization of conspiracy theories in text generation. *arXiv preprint*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7315–7330, Online. 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. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020c. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint*.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. 2023. Merge, then compress: Demystify efficient smoe with hints from its routing policy. *arXiv preprint arXiv:2310.01334*.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020a. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. 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.

Yiwei Li, G Brian Golding, and Lucian Ilie. 2020b. DELPHI: Accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics*, 37(7):896–904.

Yuhua Li, Zuhair A. Bandar, and David Mclean. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

Chao-Chun Liang, Yu-Shiang Wong, Yi-Chung Lin, and Keh-Yih Su. 2018. A meaning-based statistical English math word problem solver. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 652–662, New Orleans, Louisiana. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020b. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020a. Birds have four legs?! NumerSense: probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021a. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint*.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. Commongen: A constrained text generation challenge for generative commonsense reasoning. *Preprint*, arXiv:1911.03705.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020c. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019a. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Kevin Lin, Ben Tan, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2019b. Ropes: Reading comprehension over paragraphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021b. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint*.

Zhisheng Lin, Han Fu, Chenghao Liu, Zhuo Li, and Jianling Sun. 2024. Pemt: Multi-task correlation guided mixture-of-experts enables parameter-efficient transfer learning. *arXiv preprint arXiv:2402.15082*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Jerry Liu. 2024. LlamaIndex, a data framework for your LLM applications. `https://github.com/run-llama/llama_index`.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for GPT-3? *arXiv preprint*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020a. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint*.

Nelson F. Liu, Tony Lee, Robin Jia, and Percy Liang. 2021b. Can small and synthetic benchmarks drive modeling innovation? A retrospective study of question answering modeling approaches. *arXiv preprint*.

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

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020b. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *arXiv preprint*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020c. Multilingual denoising pre-training for neural machine translation. *arXiv preprint*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *arXiv preprint*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *arXiv preprint*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv preprint*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security*. Springer, Cham.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. Routing to the expert: Efficient reward-guided ensemble of large language models. *arXiv preprint arXiv:2311.08692*.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *arXiv preprint arXiv:2406.15479*.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, page 6309–6317.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. EventPlus: A temporal event understanding pipeline. *arXiv preprint*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011a. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011b. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. *arXiv preprint*.

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic prediction of discourse connectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Jihang Mao and Wanli Liu. 2019. A BERT-based approach for automatic humor detection and scoring. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 197–202.

Gary Marcus. 2020. The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint*.

Gary Marcus and Ernest Davis. 2020. GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Kim Marriott, Bongshin Lee, Matthew Butler, Ed Cutrell, Kirsten Ellis, Cagatay Goncu, Marti Hearst, Kathleen McCoy, and Danielle Albers Szafir. 2021. Inclusive data visualization for people with disabilities: A call to action. *Interactions*, 28(3):47–51.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Javier Marín, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Philip Massey, Patrick Xia, David Bamman, and Noah A. Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint*.

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Maxine. 2023. Llama-2, mo' lora. `https://crumbly.medium.com/llama-2-molora-f5f909434711`.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Mayne. 2020. OpenAI API alchemy: Emoji storytelling. *Andrew Mayne blog*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint*.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

Momoh Karmah Mbogba, Zeeshan Haider, S.M. Chapal Hossain, Daobin Huang, Kashan Memon, Fazil Panhwar, Zeling Lei, and Gang Zhao. 2018. The application of convolution neural network based cell segmentation during cryopreservation. *Cryobiology*, 85:95–104.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. *arXiv preprint*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint*.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *arXiv preprint*.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*.

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.

Christine Palm Meister. 2007. Phraseologie des schwedischen. In H. Burger et al., editor, *Phraseologie/Phrasology*, volume 2, pages 673–681. De Gruyter Mouton.

Francisco S. Melo, Carla Guerra, and Manuel Lopes. 2018. Interactive optimal teaching with unknown learners. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2567–2573.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Copenhagen, Denmark. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint*.

William Merrill. 2020. On the linguistic capacity of real-time counter automata. *arXiv preprint*.

Elliot Meyerson and Risto Miikkulainen. 2017. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. *ArXiv*, abs/1711.00108.

Wolfgang Mieder. 2019. "Andere zeiten, andere lehren": Sprach-und kulturgeschichtliche betrachtungen zum sprichwort. In K. Steyer, editor, *Wortverbindungen - mehr oder weniger fest*, pages 415–438. De Gruyter, Berlin.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of English puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–729, Beijing, China. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, page 25–30, Menlo Park. Association for the Advancement of Artificial Intelligence.

Republic of China Ministry of the Interior. 2018. National name statistical analysis. `https://www.ris.gov.tw/documents/data/5/2/107namestat.pdf` (Accessed 3 March 2021).

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, arXiv:2104.08773.

Ishan Misra, Abhinav Shrivastava, Abhinav Kumar Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia. Association for Computational Linguistics.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with deep reinforcement learning. *arXiv preprint*.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38, Minneapolis, Minnesota. Association for Computational Linguistics.

Alireza Mohammadshahi, Ali Shaikh, and Majid Yazdani. 2024. Routoo: Learning to route to large language models effectively. *Preprint*, arXiv:2401.13979.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2016b. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? In *Conference on Empirical Methods in Natural Language Processing*.

Karl Mulligan, Robert Frank, and Tal Linzen. 2021. Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 125–135, Online. Association for Computational Linguistics.

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. 2024. Learning to route among specialized experts for zero-shot generalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36829–36846. PMLR.

Mohammed Muqeeth, Haokun Liu, and Colin Raffel. 2023. Soft merging of experts with adaptive routing. *arXiv preprint arXiv:2306.03745*.

Yoichi Murakami and Kenji Mizuguchi. 2010. Applying the naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15):1841–1848.

Gregory L. Murphy. 1988. Comprehending complex concepts. *Cognitive Science*, 12(4):529–562.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint*.

Ramanujapuram Narasimhachar. 1988. *History of Kannada Literature: Readership Lectures*. Asian Educational Services, New Dehli.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Nam Nguyen and Yunsong Guo. 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th International Conference on Machine Learning*, page 681–688, New York, NY, USA. Association for Computing Machinery.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4885–4901, Online. Association for Computational Linguistics.

Marilyn Nippold, Melissa Allen, and Dixon Kirsch. 2001. Proverb comprehension as a function of reading proficiency in preadolescents. *Language Speech and Hearing Services in Schools*, 32:90.

Masaaki Nishino, Sho Takase, Tsutomu Hirao, and Masaaki Nagata. 2019. Generating natural anagrams: Towards language generation under hard combinatorial constraints. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6408–6412, Hong Kong, China. Association for Computational Linguistics.

David Noever, Matt Ciolino, and Josh Kalin. 2020. The chess transformer: Mastering play using generative language models.

David Nolan and Akina Mikami. 2013. "The things that we have to do": Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70.

Klaus Oberauer, Robin Hörnig, Andrea Weidenfeld, and Oliver Wilhelm. 2005. Effects of directionality in deductive reasoning, II. Premise integration and conclusion evaluation. *The Quarterly Journal of Experimental Psychology Section A*, 58(7):1225–1247.

Klaus Oberauer and Oliver Wilhelm. 2000. Effects of directionality in deductive reasoning, I. The comprehension of single relational premises. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1702–1712.

The Working Committee on the Revision of the National Standard Occupational Classification. 2015. *Standard Occupational Classification of the People's Republic of China*. China Labour and Social Security Publishing House. http://www.jiangmen.gov.cn/bmpd/jmsrlzyhshbzj/zwfw/bmjd/jdks/content/post_2334804.html (Accessed 4 June 2022).

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *Preprint*, arXiv:2406.18665.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Peter-Michael Osera and Steve Zdancewic. 2015. Type-and-example-directed program synthesis. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '15*, page 619–630, New York, NY, USA. Association for Computing Machinery.

Oleksiy Ostapenko, Zhan Su, Edoardo Maria Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas Caccia, and Alessandro Sordoni. 2024. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*.

Jahna C. Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: A preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 27–44, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Kartikey Pant and Tanvi Dadu. 2020. Sarcasm detection using context separators in online discourse. *arXiv preprint*.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2021. A review of speaker diarization: Recent advances with deep learning. *arXiv preprint*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Anthony M. Paul. 1970. Figurative language. *Philosophy & Rhetoric*, 3(4):225–248.

Ali Payani and Faramarz Fekri. 2019. Learning algorithms via neural logic networks. *arXiv preprint*.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco.

Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Devin Pelser and Hugh Murrell. 2019. Deep and dense sarcasm detection. *arXiv preprint*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint*.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! An annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint*.

Dessislava Petrova-Antonova and Rumyana Tancheva. 2020. Data cleaning: A case study with OpenRefine and Trifacta Wrangler. In *Quality of Information and Communications Technology*, pages 32–40, Cham. Springer.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–503.

Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint*.

Steve Piantadosi. 2020. Fleet system.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Tony A. Plate. 1994. *Distributed representations and nested compositional structure*. Ph.D. thesis, University of Toronto, Toronto.

Tony A. Plate. 2003. *Holographic Reduced Representations: Distributed Representation for Cognitive Structures*. CSLI, Stanford, CA.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Nadia Polikarpova, Ivan Kuraj, and Armando Solar-Lezama. 2016. Program synthesis from polymorphic refinement types. In *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '16*, page 522–538, New York, NY, USA. Association for Computing Machinery.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *arXiv preprint*.

Edoardo Maria Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. 2023. Combining parameter-efficient modules for task-level generalisation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 687–702.

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

Octavian Popescu and Carlo Strapparava. 2015. SemEval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878, Denver, Colorado. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32:17309–17320.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. *arXiv preprint arXiv:2311.11077*.

Norman M. Prentice and Robert E. Fathman. 1975. Joking riddles: A developmental index of children's humor. *Developmental Psychology*, 11:210–216.

Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, and Simon Lacoste-Julien. 2020. An analysis of the adaptation speed of causal models. *arXiv preprint*.

James Pustejovsky, Robert Ingria, Roser Saurí, José Castaño, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2004. The specification language TimeML.

Qimingtong. 2016. What are the most popular names chinese parents give their babies? a perspective from big data. `https://www.qimingtong.com/article/0` (Accessed 3 March 2021).

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Willard V.O. Quine. 1951. Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43.

Quora, Inc. 2017. Quora question pairs.

Dragomir R. Radev, Lori Levin, and Thomas E. Payne. 2008. The North American computational linguistics olympiad (NACLO). In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 87–96, Columbus, Ohio. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, page 337–346, New York, NY, USA. Association for Computing Machinery.

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. *Journal of Machine Learning Research*, 21:1–67.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Comput. Surv.*, 53(2).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Prajit Ramachandran and Quoc V. Le. 2019. Diversity and depth in per-example routing models. In *International Conference on Learning Representations*.

Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. 2022. Recycling diverse models for out-of-distribution generalization. *arXiv preprint arXiv:2212.10445*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Ian Ravenscroft. 2019. Folk psychology as a theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2019 edition. Metaphysics Research Lab, Stanford University.

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.

Scott Reed and Nando de Freitas. 2015. Neural programmer-interpreters. *arXiv preprint*.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. *arXiv preprint*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint*.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California. Association for Computational Linguistics.

He Ren and Quan Yang. 2017. Neural joke generation.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volume 1*, page 448–453, San Francisco. Morgan Kaufmann.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011a. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011b. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *AAAI Spring Symposium*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 123–150, Online. Association for Computational Linguistics.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Kenneth J. Rothman and Sander Greenland. 2005. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150.

Joshua Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for NLP. *arXiv preprint*.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Latent multi-task architecture learning. In *AAAI Conference on Artificial Intelligence*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Etta Rudolph and Alexander W. Kocurek. 2020. Comparing conventions. In *Proceedings of Semantics and Linguistic Theory*, pages 294–313, Washington, D.C. Linguistic Society of America.

Rosa Rugani, Giorgio Vallortigara, Konstantinos Priftis, and Lucia Regolin. 2015. Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, 347(6221):534–536.

Joshua S. Rule. 2020. *The child as hacker: Building more human-like models of learning*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Joshua S. Rule, Joshua B. Tenenbaum, and Steven T. Piantadosi. 2020. The child as hacker. *Trends in Cognitive Sciences*, 24(11):900–915.

D. E. Rumelhart, J. L. McClelland, and PDP Research Group, editors. 1986. *Parallel Distributed Processing. Volume 1: Foundations*. MIT Press, Cambridge, MA.

Stuart J. Russell and Peter Norvig. 2002. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? On the monolingual performance of multilingual language models. *arXiv preprint*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Meeting of the Association for Computational Linguistics (ACL)*.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A challenge on learning from small data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2016. Robsut wrod reocginiton via semi-character recurrent neural network. *arXiv preprint*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020a. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8732–8740, New York, NY, USA. Association for the Advancement of Artificial Intelligence.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020b. WINOGRANDE: An adversarial Winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI-20*, pages 8732–8734, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Suresh Kumar Sanampudi and G. Vijaya Kumari. 2010. Temporal reasoning in natural language processing: A survey. *International Journal of Computer Applications*, 1(4):53–57.

Evan Sandhaus. 2008. The New York Times annotated corpus LDC2008T19. *Linguistic Data Consortium*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon

Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. 2021. Symbolic behaviour in artificial intelligence. *arXiv preprint*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *arXiv preprint*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *arXiv preprint*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. *arXiv preprint*.

Megan Scudellari. 2017. Cryopreservation aims to engineer novel ways to freeze, store, and thaw organs. *Proceedings of the National Academy of Sciences*, 114(50):13060–13062.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017a. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017b. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint*.

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L. Dill. 2018. Learning a SAT solver from single-bit supervision. *arXiv preprint*.

Lutfi Kerem Senel and Hinrich Schütze. 2021. Does he wink or does he nod? A challenging benchmark for evaluating word understanding of language models. *arXiv preprint*.

Luzi Sennhauser and Robert C. Berwick. 2018. Evaluating the ability of LSTMs to learn context-free grammars. *arXiv preprint*.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. *arXiv preprint*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. 2014. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Usman Shahid and Elena Zheleva. 2021. Counterfactual learning in networks: An empirical study of model dependence.

Janelle Shane. 2020. All your questions answered. *AI Weirdness*.

David Elliot Shaw, William R. Swartout, and C. Cordell Green. 1975. Inferring LISP programs from examples. In *IJCAI'75: Proceedings of the 4th International Joint Conference on Artificial Intelligence*, volume 1, pages 260–267. Artificial Intelligence Laboratory, Cambridge, MA.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2016. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. 2024. Learning to decode collaboratively with multiple language models. *arXiv preprint arXiv:2403.03870*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. 2020. Neural logic reasoning. *arXiv preprint*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. 2023. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*.

Abu Awal Md Shoeb and Gerard de Melo. 2020. EmoTag1200: Understanding the association between emojis and emotions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8957–8967, Online. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint*.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2020. DiscSense: Automated semantic analysis of discourse markers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 991–999, Marseille, France. European Language Resources Association.

Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-shot recommendation as language modeling. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 223–230, Cham. Springer.

Gurdeep Singh, Kaustubh Dhole, Priyadarshini P. Pai, and Sukanta Mondal. 2014. SPRINGS: Prediction of protein-protein interaction sites using artificial neural networks. *Journal of Proteomics & Computational Biology*, 1:7.

Rishabh Singh and Sumit Gulwani. 2015. Predicting a correct program in programming by example. In *Computer Aided Verification*, pages 398–414, Cham. Springer International Publishing.

Rishabh Singh and Sumit Gulwani. 2016. Transforming spreadsheet data types using examples. In *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL '16*, page 343–356, New York, NY, USA. Association for Computing Machinery.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint*.

Natalia Skachkova, Thomas Trost, and Dietrich Klakow. 2018. Closing brackets with recurrent neural networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 232–239, Brussels, Belgium. Association for Computational Linguistics.

Douglas R. Smith. 1984. The synthesis of LISP programs from examples: A survey. In Alan W. Biermann, Gerhard Guiho, and Yves Kodratoff, editors, *Automatic Program Construction Techniques*, pages 307–324. Macmillan, New York.

Paul Smolensky. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv preprint*.

Mahmoud Soltani Firouz, Ali Farahmandi, and Soleiman Hosseinpour. 2021. Early detection of freeze damage in navel orange fruit using nondestructive low intensity ultrasound coupled with machine learning. *Food Analytical Methods*, 14:1140–1149.

Dan Sperber and Deirdre Wilson. 2002. Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1-2):3–23.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng

He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, page 2807–2813, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*, Syntax and Semantics 9, pages 315–332. Brill, Leiden.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging Evidence in Language and Communication Research 14. John Benjamins, Amsterdam.

Bernd Steinbach and Roman Kohut. 2002. Neural networks – a model of boolean functions. *5th International Workshop on Boolean Problems, Freiburg, Sept. 2002.*

Sebastian U. Stich. 2018. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint*.

Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych. 2020. Metaphoric paraphrase generation. *arXiv preprint*.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! Computing semantic relatedness using Wikipedia. In *AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence*, volume 2, page 1419–1424. Association for the Advancement of Artificial Intelligence.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017. Prerequisite skills for reading comprehension: Multi-perspective analysis of MCTest datasets and systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. 2020. Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE*, 15(8):1–39.

Ximeng Sun, Rameswar Panda, and Rogério Schmidt Feris. 2019b. Adashare: Learning what to share for efficient deep multi-task learning. *ArXiv*, abs/1911.12423.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *Advances in Neural Information Processing Systems*.

Mirac Suzgun, Yonatan Belinkov, Stuart Shieber, and Sebastian Gehrmann. 2019a. LSTM networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 44–54, Florence. Association for Computational Linguistics.

Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M. Shieber. 2019b. Memory-augmented recurrent neural networks can learn generalized Dyck languages. *arXiv preprint*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Colton Swingle, Henry Mellsop, and Alex Langshur. 2021. ChePT – applying deep neural transformer models to chess move prediction and self-commentary.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2018. Quarel: A dataset and models for answering questions about qualitative relationships. *CoRR*, abs/1811.08048.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. "quartz: An open-domain dataset of qualitative relationship questions". *EMNLP*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019a. oLMpics – on what language model pre-training captures. *arXiv preprint*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019b. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Derek Tam, Mohit Bansal, and Colin Raffel. 2023. Merging by matching models in task subspaces. *arXiv preprint arXiv:2312.04339*.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015. Learning to recommend quotes for writing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, page 2453–2459, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. 2024. Merging multi-task models via weight-ensembling mixture of experts. *Preprint*, arXiv:2402.00433.

Jan Arne Telle, José Hernández-Orallo, and Cèsar Ferri. 2019. The teaching size: Computable teachers and learners for universal languages. *Machine Learning*, 108:1653–1675.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint*.

Paul Thagard, Keith J. Holyoak, Greg Nelson, and David Gochfeld. 1990. Analog retrieval by constraint satisfaction. *Artificial Intelligence*, 46(3):259–310.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: A survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-15*, pages 1923–1929.

Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217.

Poonam B. Thorat, Rajeshwari M. Goudar, and Sunita Barve. 2015. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110:31–36.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoyu Tong. 2021. Metaphor paraphrasing and word sense disambiguation: Toward a new approach to automated metaphor.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. Learning chess blindfolded: Evaluating language models on state tracking. *arXiv preprint*.

David Toubiana, Nir Sade, Lifeng Liu, Maria del Mar Rubio Wilhelmi, Yariv Brotman, Urszula Luzarowska, John P. Vogel, and Eduardo Blumwald. 2020. Correlation-based network analysis combined with machine learning techniques highlight the role of the gaba shunt in brachypodium sylvaticum freezing tolerance. *Scientific Reports*, 10:no. 4489.

Andrew Trask, Felix Hill, Scott Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. *arXiv preprint*.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2010. Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research*, 37(1):1–40.

George Tsatsaronis, Iraklis Varlamis, Michalis Vazirgiannis, and Kjetil Nørvåg. 2009. Omiotis: A thesaurus-based measure of text relatedness. In *Machine Learning and Knowledge Discovery in Databases*, pages 742–745, Berlin. Springer.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1605–1615, Melbourne, Australia. Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. *arXiv preprint*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354.

Ellen M. Voorhees. 2002. Overview of the trec 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss. 2020. Does GPT-2 know your phone number? *Berkeley Artificial Intelligence Research blog*.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Hanqing Wang, Bowen Ping, Shuo Wang, Xu Han, Yun Chen, Zhiyuan Liu, and Maosong Sun. 2024. Lora-flow: Dynamic lora fusion for large language models in generative tasks. *arXiv preprint arXiv:2402.11455*.

Jianfeng Wang, Rong Xiao, Yandong Guo, and Lei Zhang. 2019c. Learning to count objects with few exemplar annotations. *arXiv preprint*.

Lingzhi Wang, Jing Li, Xingshan Zeng, Haisong Zhang, and Kam-Fai Wong. 2020b. Continuity of topic, interaction, and query: Learning to quote in online conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6640–6650, Online. Association for Computational Linguistics.

Po-Wei Wang, Priya L. Donti, Bryan Wilder, and Zico Kolter. 2019d. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *arXiv preprint*.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022a. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. *arXiv preprint arXiv:2205.12410*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

Ingmar Weber and Alejandro Jaimes. 2011. Who uses web search for what: And how. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, page 15–24, New York, NY, USA. Association for Computing Machinery.

David Wechsler. 2008. *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. Pearson, San Antonio.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint*.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Sarah White, Elisabeth Hill, Francesca Happé, and Uta Frith. 2009. Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, 80(4):1097–1117.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ulrike Willinger, Andreas Hergovich, Michaela Schmoeger, Matthias Deckert, Susanne Stoettner, Iris Bunda, Andrea Witting, Melanie Seidler, Reinhilde Moser, Stefanie Kacena, David Jaeckle, Benjamin Loader, Christian Mueller, and Eduard Auff. 2017. Cognitive and emotional demands of black humour processing: The role of intelligence, aggressiveness and mood. *Cognitive Processing*, 18:159–167.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Basil Blackwell, Oxford.

Thomas Wolf. 2019. Some additional experiments extending the tech report "assessing BERT's syntactic abilities" by Yoav Goldberg.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*.

Min-Sub Won, YunSeok Choi, Samuel Kim, Cheol-Won Na, and Jee-Hyong Lee. 2021. An embedding method for unseen words considering contextual information and morphological information. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing, SAC '21*, page 1055–1062, New York, NY, USA. Association for Computing Machinery.

Mitchell Wortsman, Maxwell C Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. 2021. Learning neural network subspaces. In *International Conference on Machine Learning*, pages 11217–11227. PMLR.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR.

Bo Wu, Pedro Szekely, and Craig A. Knoblock. 2012. Learning data transformation rules through examples: Preliminary results. In *Proceedings of the Ninth International Workshop on Information Integration on the Web, IIWeb '12*, New York, NY, USA. Association for Computing Machinery.

Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. 2023. pi-tuning: Transferring multimodal foundation models with optimal multi-task interpolation. In *International Conference on Machine Learning*, pages 37713–37727. PMLR.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint*.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint*.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. 2021. The causal-neural connection: Expressiveness, learnability, and inference. *arXiv preprint*.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020a. Recipes for safety in open-domain chatbots. *arXiv preprint*.

Jingwei Xu, Junyu Lai, and Yunpeng Huang. 2024. Meteora: Multiple-tasks embedded lora for large language models. *arXiv preprint arXiv:2405.13053*.

Silei Xu, Sina Semnani, Giovanni Campagna, and Monica Lam. 2020b. AutoQA: From databases to QA semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 422–434, Online. Association for Computational Linguistics.

Yang Xu, Jiawei Liu, Wei Yang, and Liusheng Huang. 2018. Incorporating latent meanings of morphological compositions to enhance word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1232–1242, Melbourne, Australia. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023a. Compeft: Compression for communicating parameter efficient updates via sparsification and quantization. *Preprint*, arXiv:2311.13171.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. 2024. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *arXiv preprint arXiv:2408.07057*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023b. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xinru Yan and Ted Pedersen. 2017. Who's to say what's funny? A computer using language models and deep learning, that's who! *arXiv preprint*.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2023. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*.

Kaiyu Yang and Jia Deng. 2019. Learning to prove theorems via interacting with proof assistants. *arXiv preprint*.

Scott Cheng-Hsin Yang and Patrick Shafto. 2017. Explainable artificial intelligence via Bayesian teaching. *Workshop on Teaching Machines, Robots, and Humans, NIPS 2017*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Qinyuan Ye, Juan Zha, and Xiang Ren. 2022. Eliciting and understanding cross-task skills with task-level mixture-of-experts. *arXiv preprint arXiv:2205.12701*.

Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 41–49, Suntec, Singapore. Association for Computational Linguistics.

Yelp, Inc. 2018. Yelp open dataset.

Yang Yi, Yih Wen-tau, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. *Association for Computational Linguistics*, page 2013–2018.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI-15*, page 1383–1389.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019a. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019b. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. *CoRR*, arXiv:2002.04326.

Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn. 2019c. Learning the Dyck language with attention-based Seq2Seq models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 138–146, Florence, Italy. Association for Computational Linguistics.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019d. ActivityNet-QA: A dataset for understanding complex web videos via question answering. *arXiv preprint*.

Eliezer Yudkowsky. 2008. Artificial intelligence as a positive and negative factor in global risk. In Nick Bostrom and Milan M. Ćirković, editors, *Global Catastrophic Risks*, pages 308–345. Oxford University Press, Oxford.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2023. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*.

Amir Zamir, Alexander Sax, Bokui (William) Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3722.

Wojciech Zaremba and Ilya Sutskever. 2014. Learning to execute. *arXiv preprint*.

Poorya Zaremoodi, Wray L. Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5810–5819, Marseille, France. European Language Resources Association.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *arXiv preprint*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019b. Defending against neural fake news. *arXiv preprint*.

Zihao Zeng, Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. 2024. Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models. *Preprint*, arXiv:2406.13233.

Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai, and Heng Tao Shen. 2020a. The gap of semantic parsing: A survey on automatic math word problem

solvers. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(09):2287–2305.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020b. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning, CHIL '20*, page 110–120, New York, NY, USA. Association for Computing Machinery.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020c. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019a. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020d. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018a. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019b. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. 2018b. Learning to count objects in natural images for visual question answering. *arXiv preprint*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019c. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint*.

Xinyu Zhao, Guoheng Sun, Ruisi Cai, Yukun Zhou, Pingzhi Li, Peihao Wang, Bowen Tan, Yexiao He, Li Chen, Yi Liang, et al. 2024a. Model-glue: Democratized llm scaling for a large model zoo in the wild. *arXiv preprint arXiv:2410.05357*.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. 2024b. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *Preprint*, arXiv:2402.09997.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. *arXiv preprint*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint*.

Jing Zhou, Zongyu Lin, Yanan Zheng, Jian Li, and Zhilin Yang. 2022. Not all tasks are born equal: Understanding zero-shot generalization. In *The Eleventh International Conference on Learning Representations*.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248, Florence, Italy. Association for Computational Linguistics.

Xiaojin Zhu. 2015. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. Sentence simplification by monolingual machine translation. In *Proceedings of the 5th International Conference on Natural Language Processing*.

Alan Zucconi. 6 Jan. 2016. The secrets of colour interpolation.

Appendix

A Extended Related Work

Model Merging. Model merging (Yadav et al., 2023b; Choshen et al., 2022; Wortsman et al., 2022; Ramé et al., 2022; Matena and Raffel, 2022; Ilharco et al., 2022; Tam et al., 2023; Jin et al., 2022; Yang et al., 2023; Zhao et al., 2024a) consolidates multiple independently trained models with identical architectures into a unified model that preserves multi-model capabilities. While simple parameter averaging suffices for models within a linearly connected low-loss parameter space (McMahan et al., 2017; Stich, 2018; Frankle et al., 2020; Wortsman et al., 2021; Li et al., 2023), more sophisticated techniques are necessary for complex scenarios. For instance, task vectors facilitate merging expert models trained on diverse domains (Ilharco et al., 2022). Additionally, methods like weighted merging using Fisher Importance Matrices (Matena and Raffel, 2022; Tam et al., 2023) and TIES-Merging, which addresses sign disagreements and redundancy (Yadav et al., 2023b) offers improved performance. As a non-adaptive expert aggregation method, merging serves as a fundamental baseline for numerous Model Editing with Regularization (MoErging) techniques.

Multitask Learning (MTL) research offers valuable insights for decentralized development. Notably, investigations into task-relatedness (Standley et al., 2020; Bingel and Søgaard, 2017; Achille et al., 2019; Vu et al., 2020; Zamir et al., 2018; Mou et al., 2016) provide guidance for designing routing mechanisms, while MTL architectures addressing the balance between shared and task-specific knowledge (Misra et al., 2016; Ruder et al., 2017; Meyerson and Miikkulainen, 2017; Zaremoodi et al., 2018; Sun et al., 2019b) offer strategies for combining expert contributions in a decentralized manner.

MoE for Multitask Learning. Recent research has extensively investigated mixture-of-experts (MoE) models for multitask learning, achieving promising results in unseen task generalization. These approaches generally fall into two categories: (1) Example Routing: Studies like Muqeeth et al. (2023); Zadouri et al. (2023); Wang et al. (2022a) train routers to dynamically select experts for each input, while Caccia et al. (2023) demonstrate the efficacy of routing at a finer granularity by splitting expert parameters into blocks. (2) Task Routing: Ponti et al. (2023) employs a trainable skill matrix to assign tasks to specific parameter-efficient modules, while Gupta et al. (2022) leverages task-specific routers selected based on domain knowledge. Ye et al. (2022) proposes a layer-wise expert selection mechanism informed by task representations derived from input embeddings. Such approaches leverage task-specific representation to allow the router to effectively select the most suitable experts for unseen tasks. While these studies differ from our setting by assuming simultaneous data access, they offer valuable insights applicable to our exploration of creating routing mechanisms over expert models.

B LLM for Task Instruction Generation.

B.1 Prompt Template

We use the following prompt with 3 randomly selected samples for each task to generate its description. The prompt is then fed into the `gpt-4-turbo` OpenAI API to get the generated task descriptions.

> *The following are three pairs of input-output examples from one task. Generate the task instruction in one sentence that is most possibly used to command a language model to produce them. In the instruction, remember to point out the skill or knowledge required for the task to guide the language model.*
>
> *- Input:*
> *- Output:*
>
> *- Input:*
> *- Output:*
>
> *- Input:*
> *- Output:*

B.2 Examples of the Generated Instructions

We provide several examples of LLM-generated instructions in this section.

WikiBio (Lebret et al., 2016a) (T0 Held-In):

- *Create a short biography using the provided facts, demonstrating knowledge in historical and biographical writing.*
- *Write a short biography based on the given factual bullet points, demonstrating profi-*

ciency in summarizing and transforming structured data into coherent narrative text.

CommonGen (Lin et al., 2020b) (T0 Held-In):

- *Generate a coherent sentence using all the given abstract concepts, requiring the skill of concept integration to form a meaningful sentence.*
- *Generate a coherent sentence by creatively combining a given set of abstract concepts.*

COPA (Huang et al., 2024b) (T0 Held-Out):

- *Identify the most logically consistent sentence from two given options based on the provided context, demonstrating reasoning and causal relationship skills.*
- *Generate the most likely outcome for a given scenario by choosing between two provided options based on contextual clues and causal reasoning.*

Date Understanding (Srivastava et al., 2023) (BigBench-Hard):

- *Calculate the date based on the given information and present it in MM/DD/YYYY format, ensuring that you accurately account for day, month, and year changes.*

Hindu Mythology Trivia (Srivastava et al., 2023) (BigBench-Lite):

- *Generate the correct answer by making use of your knowledge in Hindu mythology and culture.*

C Demonstrating Compositional Generation

In addition to significant improvements on held-in tasks, `GLIDER` demonstrates strong performance on held-out tasks, showcasing its generalization capability. To further examine this ability to handle unseen tasks by composing experts, we provide specific task examples illustrating the association between selected experts and the evaluated task. As Figure 2 shows, `GLIDER` primarily selects two experts for the COPA (T0 held-out) task, corresponding to CosmosQA and QuaRel. The following three examples from these tasks demonstrate their close semantic relationship:

- **COPA**:
  - Question: *Everyone in the class turned to stare at the student. Select the most plausible cause: - The student's phone rang. - The student took notes.*
  - Answer: *The student's phone rang.*
- **CosmosQA:**
  - Question: *That idea still weirds me out . I made a blanket for the baby 's older sister before she was born but I completely spaced that this one was on the way , caught up in my own dramas and whatnot . Luckily , I had started a few rows in white just to learn a stitch ages ago , and continuing that stitch will make an acceptable woobie , I think . According to the above context, choose the best option to answer the following question. Question: What did I make for the baby . Options: A. I made a carseat . B. None of the above choices . C. I made a crb . D. I finished a pair of booties .*
  - Answer: *D.*
- **QuaRel:**
  - Question: *Here's a short story: A piece of thread is much thinner than a tree so it is (A) less strong (B) more strong. What is the most sensical answer between "Thread" and "Tree"?*
  - Answer: *Thread.*

D Datasets and Metric

The specific details of all the datasets we use in this work are provided in this section.

D.1 T0 Held-In Datasets

- **CommonsenseQA** (Talmor et al., 2019b) under *MIT License*, evaluated by accuracy.
- **DREAM** (Sun et al., 2019a) under *MIT License*, evaluated by accuracy.
- **QUAIL** (Rogers et al., 2020) under *CC BY-SA 4.0*, evaluated by accuracy.
- **QuaRTz** (Tafjord et al., 2019) under *Apache 2.0 License*, evaluated by accuracy.
- **Social IQA** (Sap et al., 2019) under *MIT License*, evaluated by accuracy.

- **WiQA** (Tandon et al., 2019) under *Apache 2.0 License*, evaluated by accuracy.
- **Cosmos QA** (Huang et al., 2019) under *MIT License*, evaluated by accuracy.
- **QASC** (Khot et al., 2020) under *Apache 2.0 License*, evaluated by accuracy.
- **Quarel** (Tafjord et al., 2018) under *Apache 2.0 License*, evaluated by accuracy.
- **SciQ** (Johannes Welbl, 2017) under *MIT License*, evaluated by accuracy.
- **Wiki Hop** (Welbl et al., 2018) under *CC BY-SA 3.0*, evaluated by accuracy.
- **Adversarial QA** (Bartolo et al., 2020) under *Apache 2.0 License*, evaluated by F1 score.
- **Quoref** (Dasigi et al., 2019) under *Apache 2.0 License*, evaluated by F1 score.
- **DuoRC** (Saha et al., 2018) under *MIT License*, evaluated by F1 score.
- **ROPES** (Lin et al., 2019b) under *Apache 2.0 License*, evaluated by F1 score.
- **Hotpot QA** (Yang et al., 2018) under *CC BY-SA 4.0*, evaluated by exact match and F1 score.
- **Wiki QA** (Yi et al., 2015) under *MIT License*, evaluated by mean average precision (MAP) and mean reciprocal rank (MRR).
- **Common Gen** (Lin et al., 2020c) under *MIT License*, evaluated by BLEU and ROUGE scores.
- **Wiki Bio** (Lebret et al., 2016b) under *CC BY-SA 3.0*, evaluated by BLEU score.
- **Amazon** (Blitzer et al., 2007) under *Proprietary License*, evaluated by accuracy.
- **App Reviews** (Maas et al., 2011a) under *Proprietary License*, evaluated by accuracy.
- **IMDB** (Maas et al., 2011b) under *Proprietary License*, evaluated by accuracy.
- **Rotten Tomatoes** (Zhu et al., 2010) under *Proprietary License*, evaluated by accuracy.
- **Yelp** (Yelp, Inc., 2018) under *Apache 2.0 License*, evaluated by accuracy.
- **CNN Daily Mail** (Hermann et al., 2015a) under *Apache 2.0 License*, evaluated by ROUGE score.
- **Gigaword** (Graff et al., 2003) under *LDC License*, evaluated by ROUGE score.
- **MultiNews** (Fabbri et al., 2019) under *MIT License*, evaluated by ROUGE score.
- **SamSum** (Gliwa et al., 2019) under *CC BY-SA 4.0*, evaluated by ROUGE score.
- **XSum** (See et al., 2017a) under *Apache 2.0 License*, evaluated by ROUGE score.
- **AG News** (Zhang et al., 2015) under *CC BY-SA 3.0*, evaluated by accuracy.
- **DBPedia** (Auer et al., 2007) under *CC BY-SA 3.0*, evaluated by accuracy.
- **TREC** (Voorhees, 2002) under *NIST License*, evaluated by accuracy.
- **MRPC** (Dolan and Brockett, 2005) under *Apache 2.0 License*, evaluated by accuracy and F1 score.
- **PAWS** (Zhang et al., 2019c) under *Apache 2.0 License*, evaluated by accuracy and F1 score.
- **QQP** (Quora, Inc., 2017) under *Quora Terms of Service*, evaluated by accuracy and F1 score.

D.2 T0 Held-Out Datasets

Held-out Tasks

- **ANLI** (Nie et al., 2020) under *MIT License*, evaluated by accuracy.
- **CB** (de Marneffe et al., 2017) under *CC-BY-SA License*, evaluated by accuracy.
- **RTE** (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) under *Apache 2.0 License*, evaluated by accuracy.
- **WSC** (Levesque et al., 2012) under *Creative Commons License*, evaluated by accuracy.
- **Winogrande** (Sakaguchi et al., 2020a) under *Apache License 2.0*, evaluated by accuracy.

- **WiC** (Pilehvar and Camacho-Collados, 2019) under *CC BY-SA 4.0 License*, evaluated by accuracy.
- **COPA** (Roemmele et al., 2011a) under *BSD-2-Clause License*, evaluated by accuracy.
- **HellaSwag** (Zellers et al., 2019a) under *MIT License*, evaluated by accuracy.
- **Story Cloze** (Mostafazadeh et al., 2016b) under *CC-BY 4.0 License*, evaluated by accuracy.

D.3 BigBench-Hard Datasets

- **abstract_narrative_understanding** (Ghosh and Srivastava, 2021; Holyoak, 2012; Nippold et al., 2001; Tan et al., 2015; Wang et al., 2020b; Mostafazadeh et al., 2016a) under *Apache 2.0 License*, evaluated by accuracy
- **abstraction_and_reasoning_corpus** (Chollet, 2019; Brown et al., 2020; Chollet, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **anachronisms** (Otterbacher et al., 2002; Popescu and Strapparava, 2015; Llorens et al., 2015; Meng et al., 2017; Geva et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **analogical_similarity** (Plate, 2003, 1994; Thagard et al., 1990; Gentner et al., 1993) under *Apache 2.0 License*, evaluated by accuracy
- **analytic_entailment** (Hume, 1739–1740; Kant, 1781/1787; Wittgenstein, 1953; Quine, 1951; Grice and Strawson, 1956; Bolukbasi et al., 2016; Kocurek et al., 2020; Rudolph and Kocurek, 2020; Kocurek and Jerzak, 2021) under *Apache 2.0 License*, evaluated by accuracy
- **arithmetic** (Brown et al., 2020; Saxton et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **ascii_word_recognition** (Child et al., 2019; Chen et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **authorship_verification** (Bischoff et al., 2020; Koppel and Schler, 2004) under *Apache 2.0 License*, evaluated by accuracy
- **auto_categorization** under *Apache 2.0 License*, evaluated by accuracy
- **bbq_lite** (Crawford, 2017; Khashabi et al., 2020b; Li et al., 2020a) under *Apache 2.0 License*, evaluated by accuracy
- **bias_from_probabilities** (Bender et al., 2021; Abid et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **boolean_expressions** (Habernal et al., 2018; Yu et al., 2020; Dua et al., 2019; Liu et al., 2020a; Sinha et al., 2019; Wang et al., 2019b; Steinbach and Kohut, 2002; Saxton et al., 2019; Payani and Fekri, 2019; Trask et al., 2018; Selsam et al., 2018; Allamanis et al., 2016; Evans et al., 2018; Shi et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **bridging_anaphora_resolution_barqa** (Hou, 2020; Hou et al., 2013; Markert et al., 2012; Rajpurkar et al., 2016) under *Apache 2.0 License*, evaluated by accuracy
- **causal_judgment** (Gordon, 2010; Bosselut et al., 2019; Halpern, 2016; Knobe, 2003) under *Apache 2.0 License*, evaluated by accuracy
- **cause_and_effect** (Gordon, 2010) under *Apache 2.0 License*, evaluated by accuracy
- **checkmate_in_one** (Alexander, 2020; Ammanabrolu et al., 2019; Dambekodi et al., 2020; Ammanabrolu et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **chess_state_tracking** (Weston et al., 2015; Côté et al., 2018; Toshniwal et al., 2021; Alexander, 2020; Chen, 2020; Noever et al., 2020; Swingle et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **chinese_remainder_theorem** under *Apache 2.0 License*, evaluated by accuracy
- **cifar10_classification** under *Apache 2.0 License*, evaluated by accuracy
- **codenames** (Kim et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **color** (Gibson et al., 2017; Zucconi, 6 Jan. 2016) under *Apache 2.0 License*, evaluated by accuracy

- **com2sense** (Singh et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **common_morpheme** (Devlin et al., 2018; Wu et al., 2016; Won et al., 2021; Xu et al., 2018; Edmiston and Stratos, 2018; El-Kishky et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **context_definition_alignment** (Senel and Schütze, 2021; Reimers and Gurevych, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **convinceme** (Lin et al., 2021b; Levy et al., 2021; Clark et al., 2018; Maynez et al., 2020; Wang et al., 2020a; Kenton et al., 2021; Xu et al., 2020a; Tamkin et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **coqa_conversational_question_answering** (Reddy et al., 2019; Radford et al., 2019; Brown et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **crash_blossom** under *Apache 2.0 License*, evaluated by accuracy
- **crass_ai** (Schölkopf et al., 2021; Teney et al., 2020; Liang et al., 2020b; Pearl, 2000; Shahid and Zheleva, 2021; Xia et al., 2021; Priol et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **cryobiology_spanish** (Scudellari, 2017; Soltani Firouz et al., 2021; Toubiana et al., 2020; Mbogba et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **cryptonite** (Efrat et al., 2021; Raganato et al., 2017; Sakaguchi et al., 2020b; Miller and Gurevych, 2015; Miller et al., 2017; Joshi et al., 2017; Oprea and Magdy, 2020; Friedlander and Fine, 2018; Lewis et al., 2020c) under *Apache 2.0 License*, evaluated by accuracy
- **cs_algorithms** under *Apache 2.0 License*, evaluated by accuracy
- **cycled_letters** (Brown et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **dark_humor_detection** (Weller and Seppi, 2019; Fan et al., 2020; Willinger et al., 2017; Yang et al., 2015; Mihalcea and Strapparava, 2005) under *Apache 2.0 License*, evaluated by accuracy
- **date_understanding** (Vashishth et al., 2018; Chambers, 2012; Kotsakos et al., 2014; Vashishtha et al., 2020, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **disambiguation_qa** (Zhao et al., 2018; Rudinger et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **discourse_marker_prediction** (Malmi et al., 2018; Nie et al., 2019; Sileo et al., 2019, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **disfl_qa** (Gupta et al., 2021; Rajpurkar et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **diverse_social_bias** (Sheng et al., 2019; Nadeem et al., 2020; Hendrycks et al., 2020; Sap et al., 2020; Gehman et al., 2020; Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019; Liang et al., 2020a; Barocas and Selbst, 2016; Cho et al., 2019; Blodgett et al., 2020; Merity et al., 2016; Socher et al., 2013; Poria et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **dyck_languages** (Chomsky and Schützenberger, 1959; Suzgun et al., 2019b; Hao et al., 2018; Hewitt et al., 2020; Hahn, 2020; Suzgun et al., 2019a; Sennhauser and Berwick, 2018; Skachkova et al., 2018; Bhattamishra et al., 2020a; Yu et al., 2019c; Ebrahimi et al., 2020; Ackerman and Cybenko, 2020; Bhattamishra et al., 2020b) under *Apache 2.0 License*, evaluated by accuracy
- **dynamic_counting** (Suzgun et al., 2019a; Skachkova et al., 2018; Bhattamishra et al., 2020a; Suzgun et al., 2019b; Yu et al., 2019c; Ebrahimi et al., 2020; Ackerman and Cybenko, 2020; Bhattamishra et al., 2020b; Sennhauser and Berwick, 2018; Merrill, 2020; Karpathy, 2015) under *Apache 2.0 License*, evaluated by accuracy
- **elementary_math_qa** (Amini et al., 2019; Ling et al., 2017; Hendrycks et al., 2021c; Patel et al., 2021; Zhang et al., 2020a; Hendrycks et al., 2021b) under *Apache 2.0 License*, evaluated by accuracy

- **emojis_emotion_prediction** (Shoeb and de Melo, 2020; Plutchik, 1980) under *Apache 2.0 License*, evaluated by accuracy
- **empirical_judgments** (Kant, 1781/1787, 1783; Spirtes et al., 2000; Pearl, 1988; Goldberger, 1972; Rothman and Greenland, 2005; Roemmele et al., 2011b; Wang et al., 2019b; Evans et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **english_proverbs** (Gyasi Obeng, 1996; Honeck, 1997; Hrisztova-Gotthardt and Aleksa Varga, 2015) under *Apache 2.0 License*, evaluated by accuracy
- **english_russian_proverbs** (Bodrova, 2007; Gvarjalaże and Mchedlishvili, 1971; Wik) under *Apache 2.0 License*, evaluated by accuracy
- **entailed_polarity** (Karttunen, 2012) under *Apache 2.0 License*, evaluated by accuracy
- **entailed_polarity_hindi** (Karttunen, 2012) under *Apache 2.0 License*, evaluated by accuracy
- **epistemic_reasoning** (Ravenscroft, 2019; Call and Tomasello, 2008; Bugnyar et al., 2016; Stalnaker, 1978; Sperber and Wilson, 2002; Nematzadeh et al., 2018; Le et al., 2019; Jiang and de Marneffe, 2019; Ross and Pavlick, 2019; Bowman et al., 2015; de Marneffe et al., 2012; Jeretic et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **evaluating_information_essentiality** (Rajpurkar et al., 2018; Hosseini et al., 2014; Levy et al., 2017; Yin and Pei, 2015; de Marneffe et al., 2008; Zhu et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **fact_checker** (Thorne et al., 2018; Lee et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **factuality_of_summary** (Eyal et al., 2019; Wang et al., 2020a; Durmus et al., 2020; Vasilyev et al., 2020; See et al., 2017b; Hermann et al., 2015b; Narayan et al., 2018; Pagnoni et al., 2021; Kryscinski et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **fantasy_reasoning** (Wang et al., 2019b, 2018; McCann et al., 2018; Bhagavatula et al., 2019; Lourie et al., 2021; Liu et al., 2020a; Saxton et al., 2019; Clark et al., 2018; Yu et al., 2019d; Zhang et al., 2018a; Johnson et al., 2016) under *Apache 2.0 License*, evaluated by accuracy
- **few_shot_nlg** (Rastogi et al., 2020; Kale and Rastogi, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **figure_of_speech_detection** (Potamias et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **forecasting_subquestions** under *Apache 2.0 License*, evaluated by accuracy
- **gem** (Gehrmann et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **gender_inclusive_sentences_german** under *Apache 2.0 License*, evaluated by accuracy
- **gender_sensitivity_chinese** (on the Revision of the National Standard Occupational Classification, 2015; Household Management Research Center, 2021, 2020, 2019; Qimingtong, 2016; Ministry of the Interior, 2018) under *Apache 2.0 License*, evaluated by accuracy
- **gender_sensitivity_english** (Bordia and Bowman, 2019; Marcus et al., 1994; Caliskan et al., 2017; Bolukbasi et al., 2016; Rudinger et al., 2018; Lu et al., 2020; Gonen and Goldberg, 2019; Hall Maudslay et al., 2019; Fellbaum, 1998) under *Apache 2.0 License*, evaluated by accuracy
- **general_knowledge** (Shane, 2020; Dhingra et al., 2017; Rajpurkar et al., 2016, 2018; Lacker, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **geometric_shapes** (Bostock et al., 2011; Marriott et al., 2021; Boillot, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **goal_step_wikihow** (Zhang et al., 2020d) under *Apache 2.0 License*, evaluated by accuracy
- **gre_reading_comprehension** (Lai et al., 2017) under *Apache 2.0 License*, evaluated by accuracy
- **hhh_alignment** under *Apache 2.0 License*, evaluated by accuracy

- **high_low_game** under *Apache 2.0 License*, evaluated by accuracy
- **hindi_question_answering** (Brown et al., 2020; Radford et al., 2019; Jain et al., 2020; Lewis et al., 2020a; Artetxe et al., 2020; Rajpurkar et al., 2016) under *Apache 2.0 License*, evaluated by accuracy
- **hinglish_toxicity** under *Apache 2.0 License*, evaluated by accuracy
- **human_organs_senses** under *Apache 2.0 License*, evaluated by accuracy
- **hyperbaton** (Forsyth, 2014) under *Apache 2.0 License*, evaluated by accuracy
- **identify_math_theorems** (Gao et al., 2021; Black et al., 2022; Brown et al., 2020; Radford et al., 2019; Wang and Komatsuzaki, 2021) under *Apache 2.0 License*, evaluated by accuracy
- **identify_odd_metaphor** (Lakoff and Johnson, 2008; Gao et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **implicatures** (Davis, 2019; George and Mamidi, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **implicit_relations** (Cain and Oakhill, 1999; Bayat and Çetinkaya, 2020; Srivastava et al., 2016; Lin et al., 2019a; Massey et al., 2015) under *Apache 2.0 License*, evaluated by accuracy
- **intent_recognition** (Brown et al., 2020; Winata et al., 2021; Madotto et al., 2020; Coucke et al., 2018; Madotto et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **international_phonetic_alphabet_nli** (Williams et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **international_phonetic_alphabet_transliterate** (Brown et al., 2020; Liu et al., 2020c; Williams et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **intersect_geometry** (Weston et al., 2015; Agrawal et al., 2015; Trask et al., 2018; Seo et al., 2014; Hosseini et al., 2014; Polu and Sutskever, 2020; Yang and Deng, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **irony_identification** (Zhang et al., 2019b; Ghanem et al., 2020; Salas-Zárate et al., 2017) under *Apache 2.0 License*, evaluated by accuracy
- **kanji_ascii** under *Apache 2.0 License*, evaluated by accuracy
- **kannada** (Prentice and Fathman, 1975; Narasimhachar, 1988; Liu et al., 2021b; Lev et al., 2004; Lin et al., 2021a) under *Apache 2.0 License*, evaluated by accuracy
- **key_value_maps** under *Apache 2.0 License*, evaluated by accuracy
- **language_games** under *Apache 2.0 License*, evaluated by accuracy
- **linguistic_mappings** (McCoy et al., 2018, 2020; Mulligan et al., 2021; Rumelhart et al., 1986; Kirov and Cotterell, 2018; Berko, 1958; Baayen et al., 1995) under *Apache 2.0 License*, evaluated by accuracy
- **list_functions** (Rule et al., 2020; Rule, 2020; Green et al., 1974; Shaw et al., 1975; Biermann, 1978; Green, 1981; Smith, 1984; Feser et al., 2015; Osera and Zdancewic, 2015; Polikarpova et al., 2016; Cropper et al., 2020; Graves et al., 2014; Reed and de Freitas, 2015; Joulin and Mikolov, 2015; Balog et al., 2016; Bošnjak et al., 2017; Gaunt et al., 2016; Chen et al., 2019b; Kitzelmann, 2010; Flener and Schmid, 2008; Gulwani et al., 2017a; Devlin et al., 2017; Ellis et al., 2020; Cropper and Muggleton, 2016; Piantadosi, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **logical_args** under *Apache 2.0 License*, evaluated by accuracy
- **logical_fallacy_detection** (Brown et al., 2020; Hendrycks et al., 2021b; Bender et al., 2021; Wachsmuth et al., 2017; Yu et al., 2020; Copi et al., 2018; Oberauer et al., 2005; Oberauer and Wilhelm, 2000) under *Apache 2.0 License*, evaluated by accuracy
- **logical_sequence** (Saxton et al., 2019; Lin et al., 2020a; Bowman and Dahl, 2021) under *Apache 2.0 License*, evaluated by accuracy
- **long_context_integration** under *Apache 2.0 License*, evaluated by accuracy

- **mathematical_induction** (Hendrycks et al., 2021c; Patel et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **matrixshapes** under *Apache 2.0 License*, evaluated by accuracy
- **metaphor_boolean** (Lakoff and Johnson, 2008; Bizzoni and Lappin, 2018) under *Apache 2.0 License*, evaluated by accuracy
- **metaphor_understanding** (Paul, 1970; Tong et al., 2021; Radford et al., 2019; Rai and Chakraverty, 2020; Shutova, 2010; Stowe et al., 2020; Mohler et al., 2016; Shutova and Teufel, 2010; Birke and Sarkar, 2006; Zayed et al., 2020; Steen et al., 2010; Tong, 2021; Bizzoni and Lappin, 2018; Tsvetkov et al., 2014) under *Apache 2.0 License*, evaluated by accuracy
- **minute_mysteries_qa** (Sugawara et al., 2017; Dunietz et al., 2020; Kočiský et al., 2018; Mostafazadeh et al., 2016a; Frermann et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **misconceptions** (Irving et al., 2018; Atanasova et al., 2020; Boller and George, 1989) under *Apache 2.0 License*, evaluated by accuracy
- **mnist_ascii** under *Apache 2.0 License*, evaluated by accuracy
- **modified_arithmetic** (Brown et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **moral_permissibility** (Hendrycks et al., 2020; Lourie et al., 2020; Thomson, 1976) under *Apache 2.0 License*, evaluated by accuracy
- **movie_dialog_same_or_different** (Park et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **movie_recommendation** (Sileo et al., 2022; Thorat et al., 2015; Barkan and Koenigstein, 2016; Harper and Konstan, 2015) under *Apache 2.0 License*, evaluated by accuracy
- **mult_data_wrangling** (Bender et al., 2021; Tamkin et al., 2021; Singh and Gulwani, 2015; Cropper et al., 2016; Wu et al., 2012; Gulwani et al., 2015; Contreras-Ochando et al., 2018, 2020; Petrova-Antonova and Tancheva, 2020; Huynh and Mazzocchi, 2012; Kandel et al., 2011; Bhupatiraju et al., 2017; Ellis and Gulwani, 2017; Gulwani et al., 2012; Gulwani, 2011; Singh and Gulwani, 2016) under *Apache 2.0 License*, evaluated by accuracy
- **multiemo** (Kocoń et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **multistep_arithmetic** (Flanagan and Dixon, 2014) under *Apache 2.0 License*, evaluated by accuracy
- **muslim_violence_bias** (Abid et al., 2021; Bender et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **natural_instructions** (Mishra et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **navigate** (Graves et al., 2016; Henaff et al., 2016; Geva et al., 2020b; Chen et al., 2019a; Kryscinski et al., 2020; Côté et al., 2018; Luketina et al., 2019; Thawani et al., 2021; Lake and Baroni, 2017) under *Apache 2.0 License*, evaluated by accuracy
- **nonsense_words_grammar** under *Apache 2.0 License*, evaluated by accuracy
- **object_counting** (Rugani et al., 2015; Wang et al., 2019c; Zhang et al., 2018b; Brown et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **odd_one_out** (Resnik, 1995, 1999; Jiang and Conrath, 1997; Li et al., 2003; Banerjee and Pedersen, 2003; Jarmasz, 2012; Hughes and Ramage, 2007; Tsatsaronis et al., 2010, 2009; Morris and Hirst, 1991; Strube and Ponzetto, 2006; Ponzetto and Strube, 2007; Gabrilovich and Markovitch, 2007; Milne and Witten, 2008; Yeh et al., 2009; Radinsky et al., 2011; Cilibrasi and Vitanyi, 2007; Deerwester et al., 1990; Reisinger and Mooney, 2010; El-Yaniv and Yanay, 2013) under *Apache 2.0 License*, evaluated by accuracy
- **paragraph_segmentation** under *Apache 2.0 License*, evaluated by accuracy
- **parsinlu_qa** (Khashabi et al., 2020a) under *Apache 2.0 License*, evaluated by accuracy

- **penguins_in_a_table** (Herzig et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **periodic_elements** under *Apache 2.0 License*, evaluated by accuracy
- **persian_idioms** under *Apache 2.0 License*, evaluated by accuracy
- **phrase_relatedness** (Asaadi et al., 2019; Levy et al., 2015; Ein Dor et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **physical_intuition** under *Apache 2.0 License*, evaluated by accuracy
- **physics** under *Apache 2.0 License*, evaluated by accuracy
- **physics_questions** (Ling et al., 2017; Amini et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **polish_sequence_labeling** (Nguyen and Guo, 2007; Rei, 2017; Gu et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **presuppositions_as_nli** (Heim, 1983; de Marneffe et al., 2019; White et al., 2018; Jeretic et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **program_synthesis** (Gulwani et al., 2017b) under *Apache 2.0 License*, evaluated by accuracy
- **protein_interacting_sites** (Dhole et al., 2014; Singh et al., 2014; Li et al., 2020b; Murakami and Mizuguchi, 2010) under *Apache 2.0 License*, evaluated by accuracy
- **python_programming_challenge** (Allamanis et al., 2018; Alon et al., 2020; Hendrycks et al., 2021a) under *Apache 2.0 License*, evaluated by accuracy
- **qa_wikidata** (Radford et al., 2019; Kwiatkowski et al., 2019; Weber and Jaimes, 2011) under *Apache 2.0 License*, evaluated by accuracy
- **question_answer_creation** under *Apache 2.0 License*, evaluated by accuracy
- **question_selection** (Rajpurkar et al., 2016) under *Apache 2.0 License*, evaluated by accuracy
- **real_or_fake_text** (Dugan et al., 2020; Ippolito et al., 2020; Solaiman et al., 2019; Zellers et al., 2019b; Brown et al., 2020; Bakhtin et al., 2019; Sandhaus, 2008; Fan et al., 2018; Marín et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **reasoning_about_colored_objects** (Hendricks et al., 2018; Hosseini et al., 2014; Winograd, 1972; Wang et al., 2016; Jayannavar et al., 2020; Suhr et al., 2019; Thomason et al., 2015; Mitchell et al., 2010; Viethen and Dale, 2008; Gatt et al., 2009; Mitchell et al., 2013; Liang et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **rephrase** under *Apache 2.0 License*, evaluated by accuracy
- **riddle_sense** (Lin et al., 2021a; Talmor et al., 2019b) under *Apache 2.0 License*, evaluated by accuracy
- **roots_optimization_and_games** (Lample and Charton, 2019; Polu and Sutskever, 2020; Amos and Kolter, 2017; Agrawal et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **ruin_names** (Attardo, 2017; Ren and Yang, 2017; Amin and Burghardt, 2020; Annamoradnejad and Zoghi, 2020; Blinov et al., 2019; Yan and Pedersen, 2017; Frolovs, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **salient_translation_error_detection** under *Apache 2.0 License*, evaluated by accuracy
- **scientific_press_release** under *Apache 2.0 License*, evaluated by accuracy
- **self_awareness** (Yudkowsky, 2008; Chella et al., 2020; Schick et al., 2021; Kounev et al., 2017; Huttunen et al., 2017; Wallace et al., 2020; Clark and Jackson, 1994; Horowitz, 2017; Branwen, 2020; Chu et al., 2017) under *Apache 2.0 License*, evaluated by accuracy
- **self_evaluation_courtroom** (Hildebrandt, 2018; Daley, 2021; King and Cook, 2020) under *Apache 2.0 License*, evaluated by accuracy

- **self_evaluation_tutoring** (Zhang et al., 2019a; Irving et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **semantic_parsing_in_context_sparc** (Yu et al., 2019b, 2018, 2019a) under *Apache 2.0 License*, evaluated by accuracy
- **semantic_parsing_spider** (Yu et al., 2018, 2019b,a) under *Apache 2.0 License*, evaluated by accuracy
- **sentence_ambiguity** under *Apache 2.0 License*, evaluated by accuracy
- **similarities_abstraction** (Nasreddine et al., 2005; Wechsler, 2008) under *Apache 2.0 License*, evaluated by accuracy
- **simp_turing_concept** (Böhm, 1964; Sun et al., 2020; Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020; Vaswani et al., 2017; Hendrycks et al., 2021b; Xu et al., 2020b; Izacard and Grave, 2020; Zhu, 2015; Cakmak and Thomaz, 2014; Goodman and Frank, 2016; Degen et al., 2020; Khan et al., 2011; Basu and Christensen, 2013; Yang and Shafto, 2017; Melo et al., 2018; Telle et al., 2019; Chater and Vitányi, 2003; Hupkes et al., 2020; Lakretz et al., 2019; Toshniwal et al., 2021; Bender and Koller, 2020; Kühl et al., 2020; Marcus and Davis, 2020; Sinha et al., 2019; McClelland et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **simple_ethical_questions** (Hendrycks et al., 2020; Lourie et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **simple_text_editing** (Branwen, 2020; Malmi et al., 2019; Faltings et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **snarks** (Brown et al., 2020; Devlin et al., 2018; Lan et al., 2019; Liu et al., 2019; Radford et al., 2019; Annamoradnejad and Zoghi, 2020; Chen and Soo, 2018; Mao and Liu, 2019; Weller and Seppi, 2019; Khodak et al., 2017; Ghosh et al., 2020; González-Ibáñez et al., 2011; Joshi et al., 2015; McCoy et al., 2019; Kaushik et al., 2019; Gardner et al., 2020; Sennrich, 2016; Burlot et al., 2018; Naik et al., 2018; Zhang et al., 2016; Felbo et al., 2017; Pant and Dadu, 2020; Pelser and Murrell, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **social_support** (Wang and Jurgens, 2018) under *Apache 2.0 License*, evaluated by accuracy
- **social_iqa** (Sap et al., 2019; Bisk et al., 2020; Talmor et al., 2019b; Zellers et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **spelling_bee** (Ginsberg, 2014) under *Apache 2.0 License*, evaluated by accuracy
- **sports_understanding** under *Apache 2.0 License*, evaluated by accuracy
- **squad_shifts** (Miller et al., 2020; Brown et al., 2020; Rajpurkar et al., 2016; Baumgartner et al., 2020; McAuley et al., 2015) under *Apache 2.0 License*, evaluated by accuracy
- **subject_verb_agreement** (Lakretz et al., 2021b, 2019; Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Goldberg, 2019; Lakretz et al., 2021a; Wolf, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **sudoku** (Wang et al., 2019d; Russell and Norvig, 2002; Garcez and Lamb, 2020; Huang et al., 2018; Hendrycks et al., 2021c) under *Apache 2.0 License*, evaluated by accuracy
- **sufficient_information** under *Apache 2.0 License*, evaluated by accuracy
- **suicide_risk** (Gaur et al., 2019; Mohammadi et al., 2019; Matero et al., 2019; Shing et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **swahili_english_proverbs** under *Apache 2.0 License*, evaluated by accuracy
- **swedish_to_german_proverbs** (Hanzén, 2007; Korhonen, 2009; Meister, 2007; Mieder, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **taboo** (Joshi et al., 2017) under *Apache 2.0 License*, evaluated by accuracy
- **talkdown** (Wang and Potts, 2019; Mendelsohn et al., 2020; Fiske, 1993; Nolan and Mikami, 2013; Breitfeller et al., 2019; Perez Almendros et al., 2020) under *Apache 2.0 License*, evaluated by accuracy

- **temporal_sequences** (Elazar et al., 2021; Pustejovsky et al., 2004; Sanampudi and Kumari, 2010; Han et al., 2020; Ma et al., 2021; Brown et al., 2020; Petroni et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **tense** (Logeswaran et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **text_navigation_game** (Vinyals et al., 2019; Küttler et al., 2020; Kanagawa and Kaneko, 2019; Noever et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **timedial** (Qin et al., 2021; Li et al., 2017; Zhou et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **topical_chat** (Gopalakrishnan et al., 2019; Mehri and Eskenazi, 2020; Gopalakrishnan et al., 2020; Hedayatnia et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **tracking_shuffled_objects** (Liu et al., 2020b; Dong et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **training_on_test_set** under *Apache 2.0 License*, evaluated by accuracy
- **truthful_qa** (Brown et al., 2020; Sellam et al., 2020; Amodei et al., 2016; Leike et al., 2018; Kenton et al., 2021; Clark et al., 2018; Bhakthavatsalam et al., 2021; Hendrycks et al., 2021b; Khashabi et al., 2020b; Kreps et al., 2020; Maynez et al., 2020; Gabriel et al., 2020; Wang et al., 2020a; Stiennon et al., 2020; Lewis et al., 2020b; Krishna et al., 2021; Gehrmann et al., 2021; Xu et al., 2020a; Dinan et al., 2019; Tamkin et al., 2021; Bowman and Dahl, 2021) under *Apache 2.0 License*, evaluated by accuracy
- **twenty_questions** (Rajpurkar et al., 2016; Choi et al., 2018; Reddy et al., 2019; Aliannejadi et al., 2019; Clark et al., 2019; Zhang et al., 2019a) under *Apache 2.0 License*, evaluated by accuracy
- **understanding_fables** (Reimers and Gurevych, 2019; Salazar et al., 2020; Wolf et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **undo_permutation** (Pham et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **unit_conversion** (Hendrycks et al., 2021c; Geva et al., 2020a) under *Apache 2.0 License*, evaluated by accuracy
- **unit_interpretation** under *Apache 2.0 License*, evaluated by accuracy
- **unnatural_in_context_learning** (Brown et al., 2020; Kaplan et al., 2020; Henighan et al., 2020; Hernandez et al., 2021; Bahri et al., 2021; Wang et al., 2019b; Hernandez et al., 2020; Hendrycks et al., 2021c,a; Liu et al., 2021a; Zhao et al., 2021; Perez et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **unqover** (Li et al., 2020a; Caliskan et al., 2017; Rudinger et al., 2018; Zhao et al., 2018; Dev et al., 2020; Stanovsky et al., 2019; Nadeem et al., 2020; Sheng et al., 2019; Zhang et al., 2020b) under *Apache 2.0 License*, evaluated by accuracy
- **web_of_lies** under *Apache 2.0 License*, evaluated by accuracy
- **what_is_the_tao** under *Apache 2.0 License*, evaluated by accuracy
- **which_wiki_edit** under *Apache 2.0 License*, evaluated by accuracy
- **word_problems_on_sets_and_graphs** (Bency et al., 2019; Mnih et al., 2013; Russell and Norvig, 2002; Besold et al., 2017; Clark et al., 2020; Wang et al., 2018; Lacker, 2020) under *Apache 2.0 License*, evaluated by accuracy
- **word_sorting** (Grover et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **word_unscrambling** (Nishino et al., 2019; Rozner et al., 2021; Jones et al., 2020; Mays et al., 1991; Edizel et al., 2019; Sakaguchi et al., 2016; Kim et al., 2015; Xue et al., 2021a; Wu et al., 2020; Rust et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **yes_no_black_white** under *Apache 2.0 License*, evaluated by accuracy

D.4 BigBench-Lite Datasets

- **auto_debugging** (Zaremba and Sutskever, 2014) under *Apache 2.0 License*, evaluated by accuracy

- **bbq_lite_json** (Crawford, 2017; Khashabi et al., 2020b; Li et al., 2020a) under *Apache 2.0 License*, evaluated by accuracy
- **code_line_description** (Alon et al., 2018) under *Apache 2.0 License*, evaluated by accuracy
- **conceptual_combinations** (Fodor, 1975; Fodor and Pylyshyn, 1988; Smolensky, 1988; Lake et al., 2017; Lake and Murphy, 2020; Marcus, 2020; Henrich et al., 2010; Murphy, 1988) under *Apache 2.0 License*, evaluated by accuracy
- **conlang_translation** (Canfield, 2010; Şahin et al., 2020; Sennrich and Zhang, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **emoji_movie** (Cruse, 2015; Instagram Engineering, 2015; Chandra Guntuku et al., 2019; Eisner et al., 2016; Mayne, 2020; Boillot, 2019) under *Apache 2.0 License*, evaluated by accuracy
- **formal_fallacies_syllogisms_negation** (Kassner and Schütze, 2019; Talmor et al., 2019a; Betz et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **hindu_knowledge** under *Apache 2.0 License*, evaluated by accuracy
- **known_unknowns** (Liu et al., 2021c; Xiao and Wang, 2021; Shuster et al., 2021; Zhou et al., 2020; Dziri et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **language_identification** (Brown, 2014) under *Apache 2.0 License*, evaluated by accuracy
- **linguistics_puzzles** (Bozhanov and Derzhanski, 2013; Radev et al., 2008; Sennrich and Zhang, 2019; Clark et al., 2018; Şahin et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **logic_grid_puzzle** under *Apache 2.0 License*, evaluated by accuracy
- **logical_deduction** under *Apache 2.0 License*, evaluated by accuracy
- **misconceptions_russian** (Thorne et al., 2018; Lee et al., 2020) under *Apache 2.0 License*, evaluated by accuracy
- **novel_concepts** (Santoro et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **operators** (Brown et al., 2020; Kassner et al., 2020; Hendrycks et al., 2021c; Saxton et al., 2019) under *Apache 2.0 License*, evaluated by accuracy
- **parsinlu_reading_comprehension** (Khashabi et al., 2020a; Xue et al., 2021b; Rajpurkar et al., 2016) under *Apache 2.0 License*, evaluated by accuracy
- **play_dialog_same_or_different** (Park et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **repeat_copy_logic** (Graves et al., 2014) under *Apache 2.0 License*, evaluated by accuracy
- **strange_stories** (Happé, 1994; White et al., 2009) under *Apache 2.0 License*, evaluated by accuracy
- **strategyqa** (Geva et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **symbol_interpretation** (Brown et al., 2020; Johnson et al., 2016; Santoro et al., 2017; Hudson and Manning, 2019; Sennrich et al., 2015) under *Apache 2.0 License*, evaluated by accuracy
- **vitaminc_fact_verification** (Schuster et al., 2021) under *Apache 2.0 License*, evaluated by accuracy
- **winowhy** (Zhang et al., 2020c; Devlin et al., 2018; Liu et al., 2019; Kocijan et al., 2019; Rahman and Ng, 2012; Sakaguchi et al., 2020b) under *Apache 2.0 License*, evaluated by accuracy

E Efficiency Analysis

`GLIDER` introduces minimal computational overhead by requiring only two lightweight operations per LoRA layer: a single cosine similarity calculation between the query's task embedding and global routing vectors and a simple vector addition to combine this with the local routing score. With typical values for N (experts) and d_g (embedding dimension) in the hundreds, this amounts to just $(N \times d_g + d_g)$ FLOPs per layer, which is negligible compared to the base model's computation.

F Detailed Performance

We list detailed performance of all baselines for each task in Table 4, 6, 5, 7, and 7.

Method	Avg	RTE	H-Swag	COPA	WIC	Winogrande	CB	StoryCloze	ANLI-R1	ANLI-R2	ANLI-R3	WSC
Multi-Task Fine-Tuning	51.6	60.1	26.9	74.8	51.3	50.9	52.7	85.1	34.7	33	33.5	64.9
Oracle Expert	57.2	66.9	36.8	89.6	52.4	57.6	59.9	96.9	34.5	34.8	36.7	63.6
Merged Experts	45.4	55.7	25.7	61.8	50.3	53.3	45.6	63.8	33.1	33.4	33.4	43.5
LoRA Hub	46.9	52.1	27.2	75.1	50.4	50.6	33.5	84.3	33.2	34.2	33.6	41.1
Glider	57.3	64.8	29.9	91.6	50.6	60.2	69.6	96.2	36.0	34.3	38.1	58.5

Table 4: Complete results on T0 Held-Out datasets.

Expert	Avg	Boolean Expression	Causal Judgment	Date Understanding	Disambiguator QA	Formal Fallacies	Geometric Shapes	Hyperbaton
Multi-Task Fine-Tuning	34.9	49.6	55.3	35.2	55.4	51.5	10.6	50
Oracle Expert	42.2	64.4	59.5	41.7	65.1	51.7	30.1	52.7
Merged Experts	35.3	60.8	58.4	38.5	45.3	50.1	10	50
LoRA Hub	32.0	59.2	49.5	36.6	30.2	50.9	24.2	50.0
Glider	34.9	52.8	59.5	39.6	57.0	50.1	10.3	50.0

Expert	Logical Detection	Movie Recommendation	Multistep Arithmetic	Navigate	Object Counting	Penguins in a Table	Reasoning about Colored Objects	Ruin Names
Multi-Task Fine-Tuning	47.9	34.8	0	50	22.5	32.9	42	19.6
Oracle Expert	45.8	49.2	1.6	50	28.1	36.9	53.2	49.6
Merged Experts	44.3	23	0.4	50	25.4	35.6	47.5	26.6
LoRA Hub	27.5	32.0	1.2	50.0	0.0	30.9	19.3	21.2
Glider	40.8	31.2	1.6	50.0	19.7	34.2	48.4	24.3

Expert	Salient Translation Error Detection	Snarks	Sports Understanding	Temporal Sequences	Track Shuffled Objects	Web of Lies	Word Sorting
Multi-Task Fine-Tuning	27.8	46.4	50.2	16.1	17.4	51.6	0.5
Oracle Expert	27	61.3	50.9	28.2	20.1	59.2	2.8
Merged Experts	24.9	44.8	51.7	19.5	17	51.6	0.9
LoRA Hub	24.9	43.6	50.3	12.8	19.7	55.6	0.0
Glider	25.6	48.1	51.4	12.8	16.1	52.8	0.0

Table 5: BIG-bench Hard (BBH) results of different methods in T0 Held-In setting

Expert	Avg	BBQ Lite Json	Conceptual Combinations	Conlong Translation	Formal Fallacies	Hindu Knowledge
Multi-Task Fine-Tuning	36.6	40.8	44.7	26	51.5	40.6
Oracle Expert	43.5	55.3	62.1	29.8	51.6	46.3
Merged Experts	36	42.5	33	28.9	50.1	40
LoRA Hub	32.4	40.1	39.8	0.2	50.1	29.1
Glider	37.5	48.4	44.7	17.1	50.1	48.6

Expert	Known Unknowns	Linguistic Puzzles	Logic Grid Puzzle	Logical Detection	Novel Concepts	Operators
Multi-Task Fine-Tuning	47.8	0	35.9	48.1	40.6	1
Oracle Expert	65.2	0	41.7	45.4	43.8	8.6
Merged Experts	45.7	0	39.6	44.3	28.1	7.1
LoRA Hub	45.7	0.0	32.8	20.1	28.1	5.7
Glider	52.2	0.0	39.6	40.8	43.8	2.9

Expert	Play Dialog Same or Different	Repeat Copy Logic	Strange Stories	Strategy QA	Vitamin C Fact Verification	Winowhy
Multi-Task Fine-Tuning	45.8	0	47.7	52.5	54.2	44.3
Oracle Expert	63.3	0	68.4	56.1	51.1	50.5
Merged Experts	36.9	0	56.3	54.3	61.3	44.3
LoRA Hub	47.5	0.0	43.7	52.9	49.9	44.3
Glider	36.9	0.0	63.8	53.6	49.8	44.5

Table 6: BIG-bench Lite (BBL) results of different methods in T0 Held-In setting

Expert	Avg	Boolean Expression	Causal Judgment	Date Understanding	Disambiguator QA	Formal Fallacies	Geometric Shapes	Hyperbaton
Multi-Task Fine-Tuning	38.9	50	61.1	36.6	65.9	52.2	9.7	51.1
Oracle Expert	45.5	66	59.5	42.3	65.1	52.9	30.1	69.3
Merged Experts	34.6	53.6	56.8	36.9	45.7	50	12	52.2
LoRA Hub	32.8	55.2	54.2	26.8	30.2	50.0	19.5	50.0
Glider	35.3	50.8	58.4	35.2	50.4	50.5	10.3	48.2

Expert	Logical Detection	Movie Recommendation	Multistep Arithmetic	Navigate	Object Counting	Penguins in a Table	Reasoning about Colored Objects	Ruin Names
Multi-Task Fine-Tuning	49.6	32.8	0	50	35.7	39.6	56.6	19
Oracle Expert	48.8	49.2	1.6	54.6	45.7	37.6	53.5	49.6
Merged Experts	42.9	23.2	0.8	50	24.1	34.2	44.5	28.3
LoRA Hub	31.4	33.8	0.0	50.0	0.0	29.5	25.1	24.6
Glider	40.1	28.2	0.4	50.0	41.1	32.2	46.9	33.0

Expert	Salient Translation Error Detection	Snarks	Sports Understanding	Temporal Sequences	Track Shuffled Objects	Web of Lies	Word Sorting
Multi-Task Fine-Tuning	39.2	59.7	51.2	27.2	15.7	53.6	0
Oracle Expert	31.5	61.3	52.5	45.3	20.3	61.6	2.9
Merged Experts	26.4	40.9	49.9	22.4	16.3	49.6	1.2
LoRA Hub	12.4	48.1	50.3	100.0	19.1	48.8	0.0
Glider	23.8	41.4	49.6	8.8	17.2	52.0	0.1

Table 7: BIG-bench Hard (BBH) results of different methods in the FLAN setting.

Expert	Avg	BBQ Lite Json	Conceptual Combinations	Conlong Translation	Formal Fallacies	Hindu Knowledge
Multi-Task Fine-Tuning	45.4	66.9	72.8	27.9	52.2	40
Oracle Expert	46.5	61.3	62.1	36	52.9	46.3
Merged Experts	34	40.2	27.2	29.1	50	36.6
LoRA Hub	31.8	36.0	29.1	2.6	50.0	25.7
Glider	35.5	49.4	35.9	10.1	50.5	47.4

Expert	Known Unknowns	Linguistic Puzzles	Logic Grid Puzzle	Logical Detection	Novel Concepts	Operators
Multi-Task Fine-Tuning	58.7	0	42.8	49.9	37.5	13.3
Oracle Expert	65.2	0	42.5	48.6	50	12.4
Merged Experts	43.5	0	37.9	42.9	34.4	7.6
LoRA Hub	52.2	0.0	29.3	26.2	34.4	0.0
Glider	41.3	0.0	34.4	40.1	25.0	3.3

Expert	Play Dialog Same or Different	Repeat Copy Logic	Strange Stories	Strategy QA	Vitamin C Fact Verification	Winowhy
Multi-Task Fine-Tuning	44.4	0	75.9	65.7	78.5	45.3
Oracle Expert	63.3	0	74.1	56.1	66.7	53.8
Merged Experts	36.9	0	46.6	52.1	47.9	44.3
LoRA Hub	63.1	0.0	28.2	46.7	51.4	44.3
Glider	37.6	0.0	61.5	52.6	66.5	48.3

Table 8: BIG-bench Lite (BBL) results of different methods in the FLAN setting.