DIFFUSION LANGUAGE MODELS CAN PERFORM MANY TASKS WITH SCALING AND INSTRUCTION-FINETUNING

Jiasheng Ye^{1,2}*, Zaixiang Zheng^{†1}, Yu Bao¹, Lihua Qian¹, Quanquan Gu^{‡1}

¹ByteDance Research ²Fudan University

jsye23@m.fudan.edu.cn {zhengzaixiang,quanquan.qu}@bytedance.com

[†]Project Lead [‡]Corresponding Author

Code available at: https://github.com/yegcjs/DiffusionLLM

ABSTRACT

The recent surge of generative AI has been fueled by the generative power of diffusion probabilistic models and the scalable capabilities of large language models (LLMs). Despite their potential, it remains elusive whether DIFFUSION-LLM can solve general language tasks comparable to their autoregressive counterparts. This paper demonstrates that scaling masked discrete diffusion models w.r.t. data, sizes, and tasks can effectively make them strong language learners. We introduce DIFFUSION-LLMs at scale by first acquiring knowledge from massive data via masked language modeling pre-training thanks to their intrinsic connections. We then reprogram pre-trained Masked LMs into DIFFUSION-LLMs via diffusive adaptation, wherein task-specific finetuning and instruction finetuning are explored to unlock their versatility in solving general language tasks. Experiments show that scaling DIFFUSION-LLMs consistently improves performance across downstream language tasks. We further discover that instruction finetuning can elicit zero-shot and few-shot in-context learning abilities that help tackle many unseen tasks by following both natural language instructions and even visual instructions for multimodal understanding, and show promise in advanced and challenging abilities such as reasoning.

1 Introduction

Recent advances in generative modeling have led to remarkable progress in the field of generative AI. In domains of continuous signals, diffusion probabilistic models have shown great success in rendering photorealistic images (Rombach et al., 2021; Ramesh et al., 2022), immersive videos (Bar-Tal et al., 2024) and synthesizing high-quality audio (Kong et al., 2020) through iterative denoising, outperforming GANs and autoregressive (AR) models, even contributing to the surge of AI art. The story is different in the domains of discrete signals comprising symbolic sequences such as natural languages, where AR large language models (large language models or LLMs, Brown et al., 2020; OpenAI, 2023) have dominated the scene, delivering impressive generalist language abilities in language understanding and generating human-like texts, and can even follow natural language instructions to perform unseen tasks.

While many recent endeavors try unifying the generation paradigms by enabling large language models to draw (Ge et al., 2023) or speak (Zhang et al., 2023), few explore generating discrete sequences such as languages with diffusion models. We suggest that the revolutionized generative abilities of diffusion models give the promise of a strong complement to autoregressive LMs for several favorable reasons, including (1) global receptive field vs. one-sided context, and (2) non-autoregressive drafting-then-revising manner vs. restrictive unidirectional generation/autoregression. Hence, an intriguing question arises: can diffusion models speak languages well?

The key to the great success of modern LLMs lies in their scalability which fosters powerful generalist capabilities. The question of the capability of DIFFUSION-LLMs is thus in turn to ask about their scalability, which can be further boiled down into the following specific research questions regarding the three key ingredients of the success of large-scale LMs, i.e., data, model sizes, and tasks:

^{*}Work was done during Jiasheng's internship at ByteDance Research.

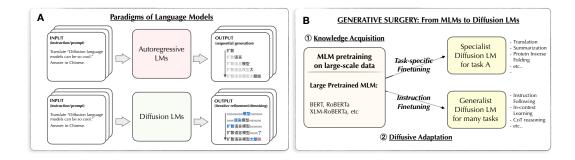


Figure 1: Overview. (A) Comparative illustration of LM paradigms, *i.e.*, autoregressive LMs *vs*. Diffusion LMs. (B) Overall illustration of the proposed DIFFUSION-LLM where large-scale pretrained masked LMs are reprogrammed to DIFFUSION-LLMs via *generative surgery*.

- (i) On scaling data. Acquiring general knowledge via self-supervised pre-training from massive unlabeled data plays a crucial role in the success of the modern NLP paradigms (Radford et al., 2018; Devlin et al., 2018), hence it is also of importance to enable Diffusion LMs to learn from massive data. Can DIFFUSION-LLMs leverage knowledge from large-scale data?
- (ii) On scaling model sizes. It has been widely observed that the larger the model size, the more competent the LMs become. Can enlarging DIFFUSION-LLMs effectively improve downstream tasks?
- (iii) On scaling tasks. What makes LLMs most attractive is they can tackle new tasks that they were never exposed to during training by following natural language and even multimodal instructions with little to no demonstrations. Can DIFFUSION-LLMs exhibit general zero-shot and few-shot in-context learning capabilities to generalize to unseen tasks?

In this paper, we delve into the potential of DIFFUSION-LLMs through the three research questions. We highlight our contributions and findings as follows:

(1) We first demonstrate the intrinsic connection between masked LMs and discrete diffusion models, which permits us to treat pre-trained masked LMs of various scales as pre-trained DIFFUSION-LLMs, without the need for expensive learning from scratch. We then reprogram pre-trained masked LMs into DIFFUSION-LLMs via diffusive adaptation, where task-specific finetuning and instruction finetuning (Wei et al., 2021) are explored for solving certain downstream tasks or general language problems, showing DIFFUSION-LLMs benefit from pre-training on large scale data. (2) We reveal that large-scale DIFFUSION-LLMs can serve as strong sequence generative models to tackle multitasks, exhibiting competitive performance compared with autoregressive LMs. And the performance consistently improves as the model sizes scale up. (3) We further elicit zero-shot and few-shot abilities for DIFFUSION-LLMs to tackle multiple unseen tasks, spanning from language and visual ones, through both language and vision instruction finetuning. Notably, DIFFUSION-LLMs demonstrate promising structured reasoning behaviors thanks to their flexible non-autoregressive generation order. Nevertheless, their capacity to tackle complex reasoning tasks remains an ongoing challenge awaiting resolution.

To sum up, we hope that our explorations provide valuable insights into the scalability of DIFFUSION-LLMs and their potential as a viable complement in tackling generative language tasks across the board.

2 Preliminaries: Diffusion Models for Sequence Generation

Language processing tasks can be unified as sequence-to-sequence problems (Raffel et al., 2020), modeling the conditional distribution $p_{\theta}(\mathbf{x}|\mathbf{c})$, where $\mathbf{x} = (\mathbf{x}^{[1]}, \mathbf{x}^{[2]}, \dots, \mathbf{x}^{[N]})$ is a target sequence composing N tokens and \mathbf{c} is the given context. For example, we may want to generate responses \mathbf{x} conditioned on the prompt \mathbf{c} , or it can be unconditional generation if no context is provided (*i.e.*, $\mathbf{c} = \phi$). As a result, one thing we care about is the capability of generative models for sequence data \mathbf{x} , *e.g.*, the prevailing autoregressive models or diffusion models. In this section, we provide the necessary background on diffusion-based sequence generative models, where we abuse the notation and use $p_{\theta}(\mathbf{x})$ for both conditional $p_{\theta}(\mathbf{x}|\mathbf{c})$ and unconditional $p_{\theta}(\mathbf{x}|\mathbf{c})$ for brevity.

Diffusion Models (Sohl-Dickstein et al., 2015) are a class of generative models characterized by a pair of Markov processes, *i.e.*, a forward diffusion process and a backward denoising process. The *forward* process $q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})$ gradually perturb the data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ into a stationary distribution $q(\mathbf{x}_T)$ with T increasingly noisy steps $\mathbf{x}_{1:T} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. The learned backward process $p_{\theta}(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$, reversely, gradually denoises the samples towards the data distribution. To fit the model $p_{\theta}(\mathbf{x}_0)$ to the data distribution $q(\mathbf{x}_0)$, the denoiser model is typically optimized by the variational bound of the negative log-likelihood (Ho et al., 2020):

$$\mathbb{E}_{q(\mathbf{x}_0)}\left[-\log p_{\theta}(\mathbf{x}_0)\right] \le \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathcal{L}_0 + \sum_{t=2}^T \mathcal{L}_t + \text{const.},\tag{1}$$

where $\mathcal{L}_0 = \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right]$, and $\mathcal{L}_t = \mathbb{E}_q \left[\text{KL}[q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) || p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \right] \right]$ for $t \in [1,T]$.

In general, diffusion models can be categorized into continuous and discrete diffusion models according to distribution type for data perturbation. Continuous diffusion models with Gaussian perturbation have demonstrated impressive performance in generating continuous signals (Rombach et al., 2021; Ho et al., 2022; Kong et al., 2020) but still struggle with satisfactory generation quality in natural languages (Li et al., 2022; Gong et al., 2022; Gao et al., 2022; Yuan et al., 2022; Ye et al., 2023). A critical challenge herein is the *pitfall of discreteness* (Ye et al., 2023) that makes Gaussian perturbation on embeddings hardly provide effective training signals. In contrast, discrete diffusion models directly operate over the discrete state space of tokens, providing an attractive alternative for generative sequence learning. Therefore in this paper, we explore developing DIFFUSION-LLMs upon discrete diffusion.

Discrete Diffusion Models (Hoogeboom et al., 2021; Austin et al., 2021) cover a subset of diffusion models for which transition probabilities between timesteps are discrete distributions. Since the forward diffusion process is applied independently to each token of a sequence \mathbf{x} , for the sake of brevity, we abuse the notation \mathbf{x}_t for arbitrary tokens at diffusion timestep t. Formally, $\mathbf{x}_t \in \{0,1\}^{|\mathcal{V}|}$ is a token represented as a one-hot vector, where \mathcal{V} is the vocabulary of all possible tokens. Let $\mathrm{Cat}(\mathbf{x};\mathbf{p})$ be a categorical distribution on \mathbf{x} with probabilities given by vector \mathbf{p} on $|\mathcal{V}|-1$ dimensional probability simplex, and the forward transition be $q(\mathbf{x}_t|\mathbf{x}_{t-1})=\mathrm{Cat}(\mathbf{x}_t;\mathbf{p}=\beta_t\mathbf{x}_{t-1}+(1-\beta_t)\mathbf{q}_{\mathrm{noise}})$, where $0 \ll \beta_t < 1$ is the noise schedule controlling the degree of perturbation at timestep t, and $\mathbf{q}_{\mathrm{noise}}$ is the probability vector of stationary distribution $q(\mathbf{x}_T)$, i.e., $q(\mathbf{x}_T)=\mathrm{Cat}(\mathbf{x}_T;\mathbf{p}=\mathbf{q}_{\mathrm{noise}})$. In this case, the distribution of corrupted sample \mathbf{x}_t given its original data \mathbf{x}_0 has a closed-form expression:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \text{Cat}(\mathbf{x}_t; \mathbf{p} = \alpha_t \mathbf{x}_0 + (1 - \alpha_t) \mathbf{q}_{\text{noise}}),$$

where $\alpha_t = \prod_{i=1}^t \beta_i$. This shows that the diffusion process is intuitively a convex combination between data and noise where the α_t controls the degree of corruption at different timesteps. In particular, α_t decreases as the timestep increases. With sufficiently large timesteps, we have $\alpha_T \approx 0$, which preserves no information from the data at the end of the diffusion process.

Different stationary distributions $\mathbf{q}_{\text{noise}}$ lead to different formulations of discrete diffusion models. One typical design is the *absorbing* diffusion with $q(\mathbf{x}_T) = \{1 \text{ if } \mathbf{x}_T = [\text{MASK}]; 0 \text{ if } \mathbf{x}_T \neq [\text{MASK}] \}$, where [MASK] is an absorbing state. According to Eq. (2), this formulation results in \mathbf{x}_t either being masked or the same as \mathbf{x}_0 , with a masking ratio $(1 - \alpha_t)$. This makes absorbing diffusion resemble masked LMs (MLM, Devlin et al., 2018) as He et al. (2023) points out.

Reparameterized Discrete Diffusion Models (RDM, Zheng et al., 2023a) reparameterize the backward transition of DIFFUSION-LLMs that reformulates the training objective of discrete diffusion models into

$$\mathcal{L}_t = \mathbb{E}\left[-\lambda_{t-1}^{(2)} \left(1 - \mathbb{1}(\mathbf{x}_t = \mathbf{x}_0)\right) \log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_t)\right],\tag{2}$$

where $\mathbb{1}(\cdot)$ is indicator function. Under the formulation of absorbing diffusion, Eqn. 2 resembles a weighted MLM objective (Devlin et al., 2018). Zheng et al. (2023a) demonstrate that Eqn. 2 is a more effective training protocol compared to Eqn. 1 for generative discrete diffusion models, showing performance on par with autoregressive LMs (Vaswani et al., 2017) on representative machine translation benchmarks for the first time. In this paper, we use RDM as our primary training objective for building our DIFFUSION-LLMs (see §A for more details).

Generative Process of Discrete Diffusion Models. Diffusion models yield new samples by their reverse generative process of iterative denoising. Under the formulation of absorbing diffusion, the

denoising process can be characterized in an iterative *mask-predict* manner (Ghazvininejad et al., 2019). Specifically, the starting sequence is initialized by all [MASK] tokens, and in each iteration, some masked tokens are replaced by the model predictions from $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ while some unmasked tokens are remasked, according to specific strategies/schedules (Ghazvininejad et al., 2019; Savinov et al., 2021; Chang et al., 2022; Zheng et al., 2023a). In this paper, we follow Zheng et al. (2023a) to unmask positions with top-k predicted $\log p_{\theta}(\mathbf{x}_0|\mathbf{x}_t)$, and mask all the rest position in each denoising step¹.

3 SCALING DIFFUSION LANGUAGE MODELS w.r.t. DATA, SIZES AND TASKS

Developing DIFFUSION-LLMs that leverage the advantages of both the generative power of both diffusion models and the scalability of large pre-trained LMs is a promising yet challenging endeavor. The key to the success of the current standard paradigm of large generative LMs is acquiring knowledge via massive pre-training and generating in a prompt-response manner for preferable output for many tasks. For DIFFUSION-LLMs, (1) how to benefit from pre-training at scale, and (2) how to best fit the prompt-response paradigm, are the crucial open questions. In this section, we will elaborate on how to empower DIFFUSION-LLMs with knowledge from pre-training of large-scale data as well as model sizes, and extend their generative capabilities for extensive downstream tasks.

3.1 KNOWLEDGE ACQUISITION VIA MLM PRE-TRAINING

The theoretical framework of discrete diffusion models has an intrinsic connection to masked language modeling (MLM), which was discussed in Austin et al. (2021); Gong et al. (2022) and He et al. (2023). Among various types of discrete diffusion models, the *absorbing* diffusion (Austin et al., 2021) resembles a *generalized* masked language modeling, which has been shown to be an effective training objective in pre-training foundation models (Devlin et al., 2018; Liu et al., 2019). Specifically, absorbing diffusion defines a stationary distribution: $q(\mathbf{x}_T) = \{1 \text{ if } \mathbf{x}_T = [\text{MASK}]; 0 \text{ if } \mathbf{x}_T \neq [\text{MASK}] \}$, where [MASK] is an absorbing token. According to Eq. (2), this formulation results in \mathbf{x}_t either being masked or the same as \mathbf{x}_0 , with a masking ratio $(1 - \alpha_t)$. Consequently, $\mathbf{x}_t = \mathbf{x}_0$ if and only if $\mathbf{x}_t \neq [\text{MASK}]$, which aligns the reparameterized training objective in Eq. (2) exactly with the masked language modeling objective.

This connection allows us to establish DIFFUSION-LLMs by pre-training with MLM objectives from massive raw textual data. We can even treat abundant community-available pre-trained MLMs (Devlin et al., 2018; Liu et al., 2019; Conneau et al., 2019) as pre-trained DIFFUSION-LLMs, and can depart from them for downstream tasks at a very low cost, bypassing the expensive pre-training stage.

3.2 DIFFUSIVE ADAPTATION: REPROGRAMMING PRE-TRAINED MLMS TO DIFFUSION-LLMS

Existing masked LMs are primarily designed to serve as sequence encoders, and are not able to generate sequences by default. Despite their connections to absorbing discrete diffusion, it is non-trivial to naively sample from masked LMs through the iterative denoising process of absorbing diffusion. One major reason is that absorbing diffusion generates sampling by iterative applying $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ from complete noise to the final prediction (i.e., ranging gradually from 100% to 0% [MASK] tokens) through different timesteps, whereas vanilla masked LMs are only pre-trained with a limited and constant masking ratio (e.g., 15%).

In order to elicit the pre-trained masked LMs' ability for sequence generation, we propose *diffusion adaptation* to eliminate the gap between pre-trained masked and DIFFUSION-LLMs, where we further finetune pre-trained Masked LMs with diffusion training objective such that sampling with the denoising process becomes possible. In particular, we follow the reparameterized training and sampling method in Zheng et al. (2023a) as described in §2. Similar to the time agnostic design in He et al. (2023), we do not introduce any extra parameters to differentiate different diffusion timesteps.

For different purposes, we perform diffusive adaptation for DIFFUSION-LLMs in two ways:

• Optimizing specialist capabilities on certain downstream tasks via task-specific finetuning. To verify the feasibility of diffusive adaptation, we finetune pre-trained masked LMs on specific datasets for each downstream task. Moreover, we further perform finetuning on pre-trained models of different scales so as to study the scalability of DIFFUSION-LLMs.

¹See §A for concrete noise schedules, and Zheng et al. (2023a) for the justification of this sampling strategy.

Table 1: SacreBLEU (Post, 2018) on IWSLT14 DE→EN and WMT14 EN→DE, and Rouge-L on Gigaword-10k. We use 10 length beams for all the results with length prediction. Results out of (inside) parentheses are obtained with length prediction (oracle target length). Results of DiffusionLM and DINOISER are quoted from Ye et al. (2023). And the results of GENIE are our reimplementation obtained with the open-source code and checkpoint of the orignal paper. "#Params.": Number of non-embedding parameters. "Type": whether the training objective and sampling method are autoregressive (AR, Vaswani et al., 2017) or follow reparameterized diffusion models (RDM, Zheng et al., 2023a). "pre-trained": whether initialized from pre-trained models. "†": The architecture is in the format of # layers / hidden dimension. For models w/ encoder, the # layers consist of layers in both encoder and decoder. "‡": For IWSLT14, we follow previous practice (Vaswani et al., 2017) and use a smaller version (39M) of Transformer-BASE, which has a dimension of 1024 in the feed-forward layers is 1024 and 4 attention heads.

	Method	Architecture [†]	#Params.	Pre-trained	IWSLT14	WMT14	Gigaword-10K	Gigaword
	AR (Vaswani et al., 2017)	12/512	39-43M [‡]	Х	33.30	26.85	10.42	34.5
	DiffusionLM (Li et al., 2022)	12/512	39-43M [‡]	X	29.11	17.41	-	-
w/ encoder	DINOISER (Ye et al., 2023)	12/512	39-43M [‡]	Х	31.44	24.62	-	-
	RDM (Zheng et al., 2023a)	12/512	39-43M [‡]	Х	32.14	26.54	16.25	-
	GENIE (Lin et al., 2022)	12/768	86M	1	-	-	-	30.24
	AR (Vaswani et al., 2017)	12/768	86M	Х	26.07	-	4.7	
w/o encoder	RDM (Zheng et al., 2023a)	12/768	86M	X	28.79 (29.12) 20	5.09 (26.86)	10.01 (10.66)	-
	DIFFUSION-LLM (RDM + XLM-R) DIFFUSION-LLM (RDM + XLM-R)		86M 9.7B	1	34.10 (35.78) 26 38.57 (40.65) 36		27.52 (28.83) 31.54 (32.57)	

• Eliciting generalist capabilities on extensive tasks via instruction finetuning. Finetuning on a collection of tasks phrased as instructions (i.e., instruction finetuning) enables LMs to better respond to instruction prompts and generalize to unseen tasks (Wei et al., 2021; Chung et al., 2022). Inspired by this, we apply diffusive adaptation to pre-trained masked LMs by instruction finetuning to study whether DIFFUSION-LLMs can acquire few-shot and zero-shot abilities like autoregressive LLMs.

Engineering considerations for scaling. Both scenarios above handle conditional sequence generation tasks from input to output, which require the model to generate target sequences according to the given prompts. Instead of incorporating an extra encoder, we handle conditional generation by organizing data in a prompt-response format². This enables our model to share the same training infrastructure as prevailing large decoder-only AR LMs (Brown et al., 2020; Touvron et al., 2023a;b; Jiang et al., 2023) while incorporating extra encoder adds to the complexity in key techniques for scaling LLMs such as pipeline parallelism (Harlap et al., 2018; Huang et al., 2019; Li & Hoefler, 2021). Our design decision ensures the scalability of our DIFFUSION-LLMs in engineering.

Besides, we incorporate a length predictor, a common practice in non-autoregressive text generation (Gu et al., 2018), to determine the lengths of predicted sequences. We pick its top-k length predictions for parallel length beam search, where k is referred to as the length beam size. During tuning, we only apply the diffusion process to the target response tokens and compute loss on them. During inference, we append the initial fully masked sequences to the prompts and denoise from them.

4 EXPERIMENTS

In this section, we first introduce our general experimental setups in $\S4$. Then we conduct three parts of experiments progressively regarding scaling on data ($\S4.2$), model sizes ($\S4.3$), and the number of tasks ($\S4.4$).

Model. Throughout the experiments, we use XLM-RoBERTa (XLM-R; Conneau et al., 2019; Goyal et al., 2021) as our foundation models, which is pre-trained on CC100 (Wenzek et al., 2020), a multilingual corpus containing 167B tokens of 100 languages, with four model sizes (numbers of non-embedding parameters) at different scales, *i.e.*, 86M, 304M, 2.8B, and 9.7B.

²A prompt-response formatted example for German→English translation (Vielen dank→Thank you): "Translate the German sentence to English. German: Vielen dank. English: Thank you."

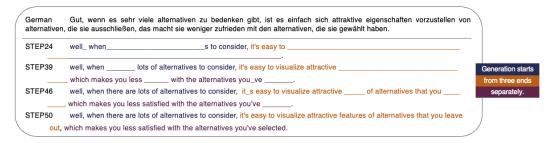


Figure 2: An exemplary generation process of DIFFUSION-LLM for machine translation. Notice that the target translation contains three segments, which are generated simultaneously by the diffusion language model.

Data & task. We investigate our approach for its specialist ability in respective downstream tasks and generalist ability to solve massive unseen tasks using natural language instructions. The datasets we use to finetune our model are as follows:

- (1) Downstream task datasets. We evaluate whether our approach can help DIFFUSION-LLMs serve as strong specialized models on multiple representative downstream tasks: (1) IWSLT14 for DE→EN translation; (2) WMT14 for EN→DE translation; (3) Gigaword-10K for text summarization; and (4) the full Gigaword summarization.
- (2) Instruction finetuning datasets. We follow Chung et al. (2022) and finetuned XLM-R models of different scales with the Flan 2022 Collection (Chung et al., 2022; Longpre et al., 2023) with diffusion training objective. It is the publicly available version of the instruction tuning data for Flan-T5 and Flan-PaLM, covering over 1.8K tasks. It combines several multitask learning datasets with instructions (Wei et al., 2021; Sanh et al., 2021; Wang et al., 2022), combined with a few extra chain-of-thought and dialog data.

4.1 EXPLORING THE DESIGN SPACE OF DIFFUSION-LLMS

We first validate that our design decisions lead to a competitive DIFFUSION-LLM, which includes (1) time-agnostic RDM and (2) no incorporation of an extra encoder. As shown in Tab. 1:

- RDM is the most performant DIFFUSION-LLM, being an ideal candidate for scaling. It outperforms both continuous diffusion language models, DiffusionLM and DINOISER. Moreover, it performs on par with AR model on IWSLT14 and WMT14, outperforming AR on Gigaword-10K, indicating that RDM is a strong enough diffusion model to exploit.
- Using an architecture without encoders has a minor effect on model performance. Despite that model without encoders slightly underperforms their counterpart with encoders, bidirectional perception of diffusion models and scaling in data can mitigate this gap, similar to the findings in Zhang et al. (2022); Patel et al. (2022). As results in Tab. 1 show, compared to AR models, RDM has a smaller performance gap between models with and without encoders. Such discrepancy becomes neglectable on large datasets like WMT14 and models initialized from the pre-trained XLM-R models.

The results verify the competence of our model at moderate scales without pre-training, preparing it to be an ideal candidate for scaling. We then build upon this to investigate its scalability *w.r.t.*, data, model sizes, and tasks.

4.2 Empowering Diffusion-LLMs with Large-scale Data

We apply diffusive adaptation on XLM-R-BASE (Conneau et al., 2019) model architecture on sequence generation benchmarks. In this way, we verify the feasibility of diffusive adaptation. Meanwhile, by comparing the performance of diffusive adapted RDMs to models trained from scratch, we confirm that our DIFFUSION-LLM takes advantage of large-scale self-supervised learning, namely the scaling with data.

pre-training at scale benefit DIFFUSION-LLMs. The results in Tab. 1 demonstrate that Training RDM through diffusive adaptation from pre-trained MLMs boosts the model performance compared to the models trained from scratch, whose impact is particularly significant on small datasets

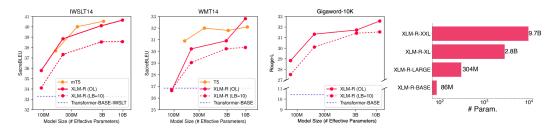


Figure 3: Scaling curves of task-specific finetuning on IWSLT14, WMT14 and Gigaword-10K. We obtain results of mT5 (Xue et al., 2020) on IWSLT14 by ourselves. The results of T5 on WMT14 are from Raffel et al. (2020). "OL": results obtained with oracle target lengths. "LB=10": length prediction results with 10 length beams. "#Params.": Number of effective parameters (*i.e.*, non-embedding parameters).

like Gigaword-10K. Compared to GENIE, a pre-trained DIFFUSION-LLM based on continuous diffusion models, RDM adapted from XLM-R is also stronger, further confirming the advantage of discrete DIFFUSION-LLM. As qualitatively shown in Fig. 2, DIFFUSION-LLMs generate fluent and semantically accurate translation³, further confirming the feasibility of our generative surgery to the pre-trained MLMs. With these findings, we confirm the feasibility of our diffusive adaptation method and verify the scalability of our DIFFUSION-LLMs regarding training data.

4.3 SCALING UP THE SIZE OF DIFFUSION-LLMS BOOSTS DOWNSTREAM TASKS

We now move on to the scalability with respect to model sizes. We finetune XLM-R models of different scales (Conneau et al., 2019; Goyal et al., 2021), whose effective parameters (*i.e.*, number of non-embedding parameters) range from <100M to 10B. Notably, when scaling up to 10B, the model shows impressive performance that surpasses base-sized models by a remarkable margin (Tab. 1).

Fig. 3 shows the scaling curve of model performance with respect to model sizes. It demonstrates that the performance of the finetuned diffusion models substantially increases as the model size increases. This shows the scaling law of DIFFUSION-LLMs in terms of model size. In addition, we also include the performance of (m) T5 (Raffel et al., 2020; Xue et al., 2020) at similar scales as references to intuitively understand how scalable our DIFFUSION-LLMs are. Note that the performance of different models is intricately affected by not only the model size but also numerous factors including model designs, pre-training budget, pre-training objectives, as well as pre-training data (Shazeer, 2020; Raffel et al., 2020; Tay et al., 2022; Scao et al., 2022; Hoffmann et al., 2022). In Fig. 3, although we see a performance gap between the finetuned (m) T5 and XLM-R models at similar scales, the discrepancy is minor and does not seem amplified as models scale up. Therefore, while there is still ample room for improving large-scale pre-trained DIFFUSION-LLMs, we believe that the path of scaling up these models holds great promise.

4.4 Instruction-Finetuning Helps Generalize to Unseen Tasks

A fascinating property that motivates scaling LMs up is that LLMs can follow instructions and show impressive few-shot or even zero-shot performance (Wei et al., 2021). We now investigate whether diffusion models can also exhibit zero-shot and few-shot performance when being scaled up.

4.4.1 Instruction finetuning elicits scalable zero-shot performance

Strict zero-shot evaluation on IWSLT14 DE→EN. We first conduct a strict zero-shot evaluation to study if DIFFUSION-LLMs can acquire zero-shot capabilities through instruction finetuning. Specifically, we

Table 2: **Zero-shot** SacreBLEU of instruction-tuned DIFFUSION-LLMs on IWSLT14 DE→EN translation. For Flan 2021, we explicitly remove all German data for strict evaluation. Results are obtained with oracle length.

Architecture	Strict Flan'21	Flan'22			
instruction-tuned Diffusio	n-LLM:				
XLM-R-BASE (85M)	7.15	21.26			
XLM-R-LARGE (304M)	22.52	25.24			
XLM-R-XL (2.8B)	27.27	28.13			
XLM-R-XXL (9.7B)	28.74	29.59			
ref: supervised AR on 160k DE→EN data: 33.30					

³The intermediate steps demonstrate that the models generate three clauses simultaneously, implying a global perception that plans the generation of the whole sequence. We consider this benefits the model on more complex generation tasks, which we discuss in §4.5.2.

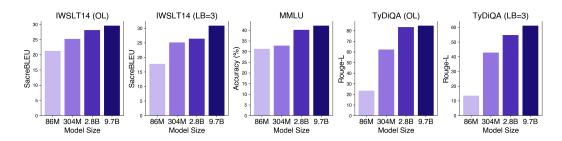


Figure 4: Zero-shot performance of instruction-tuned DIFFUSION-LLMs (Flan-XLM-Rs) of different scales. OL means the results are obtained with oracle length, while LB means the number of length beams to sample the target with length prediction. The model sizes refer to the number of non-embedding parameters.

evaluate on IWSLT14 DE→EN translation task, for which we instruction-finetune DIFFUSION-LLMs on Flan 2021 Collection (Wei et al., 2021) with all German data removed to ensure that the DE→EN translation becomes a strictly unseen task. As shown in Tab. 2, the instruction-tuned DIFFUSION-LLMs demonstrate scalable zero-shot performance even without finetuning with German data, signifying that DIFFUSION-LLMs are able to follow instructions.

Extensive zero-shot evaluation with large-scale instruction tuning. We then follow the recommended settings and conduct larger-scale instructions tuning on the full Flan 2022 Collection (Longpre et al., 2023) and run extensive evaluations⁴. Following Chung et al. (2022), we named our instruction-tuned checkpoints on Flan 2022 Collection as Flan-XLM-R. The results in Fig. 4 suggest that the Flan-XLM-R models are indeed general-purpose zero-shot learners, and their zero-shot performance substantially improves as the model scales. In particular, we highlight the results on IWSLT14. The largest model, Flan-XLM-R-XXL even achieves a 30.90 zero-shot ScareBLEU score, only 2.4 below the performance of widely adopted supervised transformer baselines (33.30 as shown in Tab. 2). This indicates the Flan-XLM-R models produce a very good language generation quality.

4.4.2 DIFFUSION-LLMS CAN LEARN IN CONTEXT

We also evaluate the in-context learning ability of the DIFFUSION-LLMs. We construct an experiment on IWSLT14 with removed instructions. In this case, the model can only rely on the demonstration to figure out the task objective.

Table 3: SacreBLEU of instruction-tuned DIFFUSION-LLMs on IWSLT14 DE→EN under oracle lengths with removed instruction, where the model can only figure out the objective of the task through demonstrations.

DIFFUSION-LLM Arch.	w/o demonstration	w/ demonstration
XLM-R-LARGE (304M)	2.58	5.38
XLM-R-XL (2.8B)	4.42	12.18
XLM-R-XXL (9.7B)	4.16	19.16

The result supports that DIFFUSION-

LLMs can learn in context. The model is unable to produce the desired outcome without prior knowledge of the task. However, when given a demonstration, it can learn to treat the task as a translation task, showing obvious performance improvement, which also scales with model sizes.

4.5 EXPLORING REASONING WITH DIFFUSION-LLMS

We are also interested in exploring the reasoning abilities of our DIFFUSION-LLMs as it is a crucial emergent ability that distinguishes LLMs from the small ones (Wei et al., 2022; Fu et al., 2023). Understanding how these models develop reasoning capabilities could provide insights into their scalability, generalization, and potential applications in complex problem-solving tasks. Moreover, investigating their reasoning mechanisms may help bridge the gap between traditional autoregressive LLMs and diffusion-based architectures, shedding light on their respective strengths and limitations in various domains. In this section, we will highlight our key findings and include detailed discussions.

⁴We continue to evaluate on the **IWSLT14** dataset. Besides, we also evaluate several datasets used in Chung et al. (2022). In detail, **MMLU** (Hendrycks et al., 2020) includes multiple-choice exam questions from 57 tasks covering humanities, social science, STEM, and more. **TyDiQA** (Clark et al., 2020) is an open-book question-answering benchmark across 8 typologically diverse languages

4.5.1 QUANTITATIVE RESULTS

As shown in Fig. 5, we find, by simply instruction tuning, even Flan-XLM-R-XXL fails to emerge non-trivial reasoning performance on GSM8K (Cobbe et al., 2021), a benchmark dataset for mathematical reasoning, and its German translated version in MGSM (Shi et al., 2022).

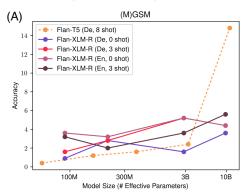
We conduct further analyses to understand the unsatisfying performance and conclude that this is due to the limitation of the pre-trained recipe instead of the DIFFUSION-LLM paradigm.

- DIFFUSION-LLM can perform reasoning tasks. Given that task-specific fine-tuning offers an effective strategy to predict the existence of emergent ability (Snell et al., 2024), we finetune Flan-XLM-R-XXL on the training set of GSM8K. We find that the model performance rockets to a non-trivial accuracy of 26.73%, confirming the capability of DIFFUSION-LLMs in reasoning.
- Reasoning performance of DIFFUSION-LLMs can be improved with better pre-training recipes. We suggest that the reasoning performance of Flan-XLM-R-XXL is limited by its pre-training recipe, which significantly falls short of modern designs (Warner et al., 2024). As a preliminary verification, we adapt an LLaMa3.1 8B into DIFFUSION-LLMs by fine-tuning on Red-Pajama⁵ and obtain 13.48 after instruction tuning on Flan 2022 and 42.77⁶ after task-specific fine-tuning on GSM8K. Both greatly outperform XLM-R-XXL with the same fine-tuning settings.

These findings support that DIFFUSION-LLMs are also able to perform reasoning as autoregressive LMs do while we need an improved pre-trained models to fully unveil the potential.

4.5.2 QUALITATIVE ANALYSIS

In reasoning tasks, a model needs to generate intermediate reasoning steps to approach the final answers, where the model heavily relies on



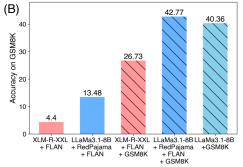


Figure 5: Evaluating reasoning on GSM datasets. (A) Performance of DIFFUSION-LLM (Flan-XLM-R) and autoregressive LLMs (Flan-T5) of different scales. Although Flan-XLM-R-XXL fails to emerge with nontrivial performance, the performance improves with (B) task-specific tuning on GSM8K training set or better foundation model (adapted from LLaMa3.1 8B Instruct by continual training on RedPajama). We also include the result of finetuning autoregressive LLaMa3.1 8B on GSM8k for reference.

the intermediate results generated by itself to predict the final answer. This leads to constraints on the generation order when performing reasoning tasks. Given the non-autoregressive nature of DIFFUSION-LLMs, we wonder whether they show different behaviors in generation orders compared to fixed left-to-right autoregressive reasoning.

Understanding target dependencies with causal graphs. Fig. 6(a) depicts the causal graph for the exemplary problem and its solution shown in Fig. 6(b). We argue that to solve the task with reasoning,

⁵We adapt the autoregressive LLaMa into a DIFFUSION-LLM by replacing the causal attentions with bidirectional ones and continue pre-train the model with diffusion objective. Meanwhile, we also post-process the output logits by right shifting for one token to fill the gap that the output of autoregressive models predicts the next token while the output of DIFFUSION-LLMs predicts the token at the same positions as the mask tokens (Gong et al., 2024).

⁶This performance is comparable to fine-tuning autoregressive LLaMa 3.1 with GSM8K training set, which is 40.36. Although we note that fine-tuning on GSM8K actually degrades the performance LLaMa 3.1 8B from near 80, we consider the fine-tuning results can represent model performance with sufficient training under the training distribution and implies similar capabilities between autoregressive and DIFFUSION-LLMs.

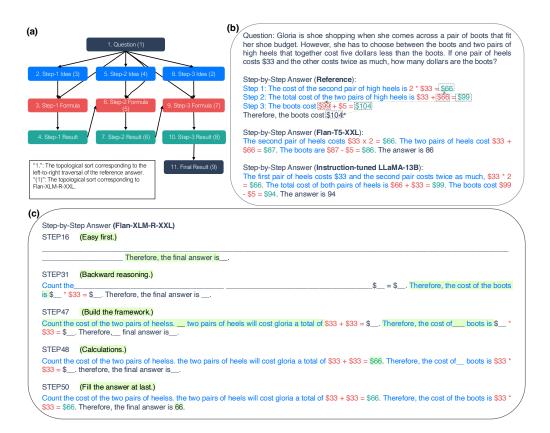


Figure 6: Qualitative investigation into the reasoning abilities of diffusion language models. (a) A causal graph (Pearl, 1998) that represents the dependencies between the reasoning steps. (b) An example question, its reference answer, and answers from autoregressive models. (c) The answer from DIFFUSION-LLM (Flan-XLM-R-XXL) and its generation process.

language models must generate tokens in an order that conforms to a *topological sort* of the causal graph. Specifically, it means the following requirements for the generation order: (1) the final results should come after the last intermediate result; (2) the intermediate results should come after listing the corresponding equation; (3) to correctly list an equation, models need to have the idea for this equation, copying calculation results from previous steps or numbers provided by the question; and (4) before these, models need to propose the idea for each step first.

DIFFUSION-LLMs can figure out feasible topological sorts on the causal graph. A follow-up question is whether the generation process of autoregressive models and our diffusion language models conform to possible topological sorts. One feasible topological sort is exactly the left-to-right traversal on the chain-of-thought text and is implicitly provided to autoregressive models during training. Diffusion language models, on the other hand, learn without a fixed generation order due to random masking. Fig. 6(c) demonstrate its generation process of solving the exemplary question. Despite incorrect final answers, the generative process does conform to a topological sort of the causal graph in Fig. 6(a). The model generates the ideas first, then writes the formulas, and finally calculates the answers. We randomly sampled 30 samples generated by DIFFUSION-LLMs and found that 21 out of these samples conformed to a topological order. This implies that diffusion language models learn to figure out feasible topological sorts, namely a structure reasoning ability.

DIFFUSION-LLMs reason with a flexible mind. Notably, diffusion language models are able to explore different topological sorts different from autoregressive models thanks to less constrained generative orders. We highlight some of the interesting patterns resulting from this.

- *Easy first.* Fig. 6(c) shows that the model fills up the fixed pattern (*i.e.*, "the final answer is") at first, showing a quite smart easy-to-hard generation behavior.
- *Planning ahead.* In Fig. 6(c), the model constructs the framework for the solution before diving into arithmetic. Actually, we have seen similar behavior in Fig. 2 where the model generates

three clauses simultaneously. Both cases demonstrate the models' global perception which helps plan the generation of the whole sequence.

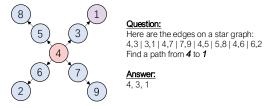
- Forward and backward reasoning. During the reasoning process in Fig. 2, on STEP 31, the model begins the solution with the idea for the last reasoning step. This shows backward reasoning behavior, a very common human behavior that is especially helpful for challenging reasoning activities such as finding mathematical proofs (Kazemi et al., 2022).
- Backtracing. The backward transition of diffusion models formally supports backtracing by remasking. In Fig. 6(c), STEP 47 erases a "the" token. This ability helps avoid accumulating errors in predicted tokens (Arora et al., 2022).

DIFFUSION-LLMS SHOW SUPERIORITY WITH REASONING THAT NESSITATES IMPLICIT PLANNING

Inspired by the qualitative observations, we consider that DIFFUSION-LLMs could be helpful when the logical reasoning process differs from the sequential (i.e., left-to-right) order of words in the text. This is common, especially for challenging reasoning tasks, implicit thought processes are required to plan before giving the outcome rationales, which is also known as meta-CoT (Xiang et al., 2025).

Path-Finding on Path-Star Graphs. To verify this, we experiment to see whether DIFFUSION-LLMs are able to solve Path-Finding on Path-Star Graphs (Bachmann & Nagarajan), a simple planning-related problem that autoregressive models struggle with⁷. As shown in Fig. 7, a path-star graph contains a central start node with multiple paths of equal length radiating outward from it, where one of the paths leads to the designated goal node. The task for the model herein is to find the correct path from start to goal. The task is simple only if the model can look ahead

Comparison with AR-LMs. The comparison between autoregressive models and DIFFUSION-LLMs in Fig. 8 reveals a significant advantage of DIFFUSION-LLMs in solving reasoning tasks that require implicit planning. To investigate this further, we fine-tune two representative models—LLaMa3.1-8B-Instruct and Flan-XLM-R-XXL—on the Path-Finding on Path-Star Graphs tasks and evaluate their generalization performance on unseen graph structures. Our evaluation encompasses two different experimental settings designed to test the models' ability to perform multi-step reasoning and implicit planning. In both cases, the autoregressive models exhibit limited success, struggling to capture the underlying graph structure and plan effectively. Their performance closely resembles a random walk, indicating an inability Flan-XLM-R-XXL.



(Degree = 4; Length = 3)

Figure 7: An example of Path-Finding on Path-Star Graph (Bachmann & Nagarajan) with degree of central start as 4 and length of each path is 3. The task inputs the graph as well as the node indices of the central start (Node 4 in the example) and goal node (Node 1 in the example), and requires a model to figure out the path from the central start and the goal node ("Node 4, Node 3, Node 1" in the example). This task serves as a minimal planning task as it can be simply solved only if the models look ahead to see which first step leads to the goal.

to discover which first step leads to the goal. The straightforward failure mode makes it representative of studying the planning capability of the models.

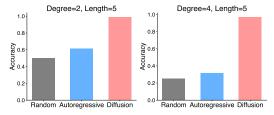
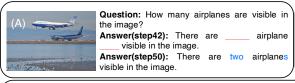
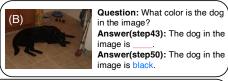


Figure 8: Comparison between autoregressive and DIFFUSION-LLMs on Path-Finding on Path-Star Graph (Bachmann & Nagarajan). experiment on two settings where the degree of the central start is 2 and 4, respectively. "Random": a random walk predictor that randomly selects a path on the star graph; "Autoregressive": fintuned LLaMa3.1-8B-Instruct; "Diffusion": DIFFUSION-LLM fine-tuned from

to leverage structural information for accurate predictions. In contrast, DIFFUSION-LLMs consis-

⁷We refer readers to Ye et al. (2024) which elaborate on the planning capability of DIFFUSION-LLM in more details.







Question: What types of animals are displayed behind the glass in the museum?

Answer(step43): In the museum, there are _____ animals, such as ____ and ___ displayed behind the glass.

Answer(step50): In the museum, there are various animals, such as zebras and giraffe, displayed behind the glass.

Figure 9: Visual question answering with DIFFUSION-LLMs. We set the iterations steps for the DIFFUSION-LLMs to generate final output as 50. Similar to the discussion in Sec. 4.5.2, the model first generates easy content and fills the answer at the end.

tently produce near-perfect predictions, demonstrating a remarkable capacity for handling implicit planning tasks.

This performance disparity highlights a key architectural difference: the bidirectional receptive field in DIFFUSION-LLMs allows the model to capture global dependencies across the input more effectively. This not only facilitates better reasoning in structured environments but also gives DIFFUSION-LLMs a clear advantage in tasks where planning and multi-step inference are required. These findings suggest that DIFFUSION-LLMs are better equipped to model complex, non-sequential dependencies, offering new possibilities for reasoning-based applications beyond the capabilities of conventional autoregressive models.

4.6 DIFFUSION-LLMS AS MULTIMODAL LEARNERS

Recent advances in large language models extend beyond language processing, aiming to unify multiple modalities end to end for seamless multimodal interactions (Zhan et al., 2024). Given the dominance of diffusion models in generating continuous signals Dhariwal & Nichol (2021); Bar-Tal et al. (2024) and excellent language capabilities shown in this study, we believe DIFFUSION-LLMs contribute to a promising paradigm for developing unified multimodal models. This motivates us to explore DIFFUSION-LLMs for multimodal tasks.

In particular, we investigate whether DIFFUSION-LLMs can tackle visual question answering (VQA). We follow LLaVa (Liu et al., 2024) to conduct two-phase training upon FLAN-XLM-R-XXL. In the first stage, we freeze the language model backbone and train a projector to map vision feature extracted from pre-trained CLIP visual encoder ViT-L/14 (Radford et al., 2021) to embeddings using the 558k subset of the LAION-CC-SBU (Liu et al., 2024). We then jointly tune the LM backbone and projectors for VQA with LLaVA-v1.5-mix665k data (Liu et al., 2024).

Tab. 4 shows the zero-shot performance of our models on the dev set of GQA (Hudson & Manning, 2019). For reference, we accordingly augment a Flan-T5-XXL model with vision understanding capability with the same recipe. The results shows meaningful performance that supports the vision understanding capability of our model, which is close to that adapted from Flan-

Table 4: Zero-shot exact match performance of DIFFUSION-LLM and AR-LLMs on the dev set of GQA (Hudson & Manning, 2019).

Model	Exact Match
DIFFUSION-LLM (Flan-XLM-R-XXL)	39.93 44.71
AR-LLM (Flan-T5-XXL)	44./1

T5. Case study on Fig. 9 shows that the model demonstrate similar behavior to what we find in the qualitative study for reasoning tasks (Sec. 4.5.2) when generating answers for vision tasks. In the three cases, the models answer in an easy-first order where they first generate content that can be copied from the question to build up the framework of the sentence and then fill in the key answers at the end. In Fig. 9(C), where the key answer contains multiple entities, DIFFUSION-LLMs fill them simultaneously, showing the capabilities to parallelly process different vision information.

These results demonstrate DIFFUSION-LLMs can also understand multimodal information similar to recent autoregressive LMs. Together with the generation capabilities, for both well-tested vision generation (Dhariwal & Nichol, 2021; Rombach et al., 2022; Bar-Tal et al., 2024) as well as language

generation abilities verified in our study, diffusion models shed light on an appealing unified paradigm for multimodal foundation models.

5 DISCUSSIONS

In this work, we pioneer the study of the scalability of DIFFUSION-LLMs to catch up with the recent advances of LLMs and facilitate the exploration of their potential. Our experiments verify their scalability regarding data, model sizes, and tasks. Further, we showcase positive prospects about their reasoning capabilities such as casual-order generation and implicit planning for further exploitation.

Latest advancement on diffusion language models. After the first release of our study of DIFFUSION-LLM, diffusion language models have attracted broad attention and plenty of closely related studies have emerged. As such, we would like to highlight the latest progress to facilitate more upcoming advancements in this field.

- Foundation. To formulate a diffusion language model, Ou et al. (2024); Sahoo et al. (2024); Shi et al. (2024); Wang et al. (2024a;b) also study masked language modeling-like objectives similar to ours and validate their effectiveness. Alternatively, Lou et al. explores learning discrete diffusion language models by learning probability ratios as the extension of score matching and Gat et al. (2025) extends the flow matching formulation. All these studies confirm the practicality of building capable diffusion language models to serve as an alternative paradigm to autoregressive language models, with specific manners evolving.
- Scaling validation. The scalability of diffusion language models under masked language modeling-like objectives has rapid progress. Gong et al. (2024) successfully build large-scale diffusion language models by adapting from autoregressive language models, offering another promising routine to gain large diffusion language models with relatively low cost. Nie et al. (2024) investigate the pretraining of diffusion language models from scratch and show a scaling law parallel to autoregressive language models, indicating similar scaling trends of the two paradigms. Nie et al. (2025) further scales up the pretrained diffusion language models to 8B parameters and 2.3T pre-training tokens with up-to-date recipes, with results highlighting the competitiveness of diffusion language models to frontier open-source autoregressive models on well-recognized benchmarks for large language models and showcase a helpful chatbot built upon large diffusion language models. Besides natural language, Wang et al. (2024a) scales diffusion language models to empower generative modeling of proteins.
- Capabilities and applications. Diffusion language models have a bidirectional receptive field and can perform refinement by nature. For this reason, recent progress has confirmed that diffusion language models show superiority in scenarios where left-to-right generation order is suboptimal. For instance, Ye et al. (2024) shows their advantage in tasks requiring implicit planning and Nie et al. (2024; 2025) shows diffusion language models can fix the reversal curse of autoregressive models (Berglund et al., 2023). Notably, the applications of diffusion language models demonstrate significant advances in scientific domains. With diffusion language models, DPLM (Wang et al., 2024a) build frontier foundation models for protein and DPLM-2 (Wang et al., 2024b) further investigate multimodal diffusion language models to unify generative modeling of protein sequences and structures.

We hope that our findings as well as our discussion on the latest progress can fuel the success of diffusion models in broader domains and also encourage investment into this compelling complement to autoregressive LLMs, which might push forward the boundary of techniques to pursue more advanced machine intelligence.

REFERENCES

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. *arXiv* preprint arXiv:2204.01171, 2022.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems, 34:17981–17993, 2021.

- Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *Forty-first International Conference on Machine Learning*.
- Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470, 2020.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* preprint arXiv:1810.04805, 2018.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*, 2023.
- Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Difformer: Empowering diffusion model on embedding space for text generation. *arXiv* preprint arXiv:2212.09412, 2022.

- Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2025.
- Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. 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)*, pp. 6112–6121, 2019.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.
- Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.
- Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. *arXiv* preprint arXiv:2105.00572, 2021.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018.
- Aaron Harlap, Deepak Narayanan, Amar Phanishayee, Vivek Seshadri, Nikhil Devanur, Greg Ganger, and Phil Gibbons. Pipedream: Fast and efficient pipeline parallel dnn training. arXiv preprint arXiv:1806.03377, 2018.
- Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967flab10179ca4b-Paper.pdf.
- Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.
- Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.
- Seyed Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. Lambada: Backward chaining for automated reasoning in natural language. *arXiv preprint arXiv:2212.13894*, 2022.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1412.6980.
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.
- Shigang Li and Torsten Hoefler. Chimera: efficiently training large-scale neural networks with bidirectional pipelines. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–14, 2021.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217, 2022.
- Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Weizhu Chen, and Nan Duan. Genie: Large scale pre-training for text generation with diffusion model. *arXiv* preprint *arXiv*:2212.11685, 2022.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv* preprint arXiv:2301.13688, 2023.
- Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.
- Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. *arXiv preprint arXiv:2410.18514*, 2024.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.
- OpenAI. Gpt-4 technical report, 2023.
- Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv* preprint arXiv:2406.03736, 2024.
- Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*, 2022.
- Judea Pearl. Graphical models for probabilistic and causal reasoning. *Quantified representation of uncertainty and imprecision*, pp. 367–389, 1998.
- Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018.

- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.
- Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. arXiv preprint arXiv:2406.07524, 2024.
- 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 Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal 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 Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2021.
- Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*, 2021.
- Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022.
- Noam Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.
- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.
- Charlie Snell, Eric Wallace, Dan Klein, and Sergey Levine. Predicting emergent capabilities by finetuning. *arXiv preprint arXiv:2411.16035*, 2024.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.
- Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024a.
- Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024b.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663, 2024.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. Cenet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4003–4012, 2020.
- Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*, 2025.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv* preprint arXiv:2010.11934, 2020.
- Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. *arXiv* preprint arXiv:2410.14157, 2024.
- Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises. *arXiv preprint arXiv:2302.10025*, 2023.
- Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*, 2022.
- Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. arXiv preprint arXiv:2402.12226, 2024.

- Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, and Orhan Firat. Examining scaling and transfer of language model architectures for machine translation. In *International Conference on Machine Learning*, pp. 26176–26192. PMLR, 2022.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv* preprint arXiv:2305.11000, 2023.
- Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023a.
- Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed language models are protein designers. In *International Conference of Machine Learning (ICML)*, 2023b.

A IMPLEMENTATION DETAILS

A.1 MODEL

Throughout this work, we mainly follow Zheng et al. (2023a) to train and sample from our diffusion language models. Specifically, we set $\lambda_{t-1}^{(2)} = 1 - \frac{t-1}{T}$ in the training objective where t is the current timestep and T is the number of total timesteps which is 50 in our experiments. Additionally, we apply label smoothing with a factor of 0.1 when we train a model without pretraining. During sampling, we also follow Ghazvininejad et al. (2019); Savinov et al. (2021); Zheng et al. (2023b) and denoise tokens with high scores in each step instead of naively sampling from the Bernoulli distributions. We use the same cosine schedule as in Zheng et al. (2023a) to decide the number of denoised tokens in each step $k = \lfloor N \cdot \cos \frac{\pi t}{2T} \rfloor$, where N is the sequence length. For full details, we refer readers to the pseudocode in the original paper (Zheng et al., 2023a, Algorithm 2). For length prediction, we feed model outputs into a one-layer transformer, apply mean pooling to the features and feed the pooled feature into an MLP classifier head. For task-specific finetuning, we remove both input and output embeddings of the tokens that do not appear in the training set.

A.2 DATA

For IWSLT14 and WMT14 machine translation tasks, we download and preprocess data following the example scripts provided by Fairseq⁸, and we use SacreBleu (Post, 2018) for evaluation⁹. And we download Gigaword-10K data from the repository of LGEB¹⁰. For (M)GSM, we follow the instruction¹¹ in the official repository of Shi et al. (2022) to process the data and prompts. Besides, we obtain the preprocessed Flan 2021¹², Flan 2022¹³, MMLU¹⁴, and TydiQA¹⁵ from shared datasets on HuggingFace¹⁶. During training with Flan 2022, we follow the recommended ratios in Chung et al. (2022) to sample training data from different subsets. We follow Chung et al. (2022) to report the MMLU performance on the validation set and adopt the GoldP setting for TyDiQA as in Chowdhery et al. (2022); Chung et al. (2022). On the few-shot settings, we randomly select demonstrations. We will also release our code and data for better reproducibility.

A.3 TRAINING DETAILS

We use Adam optimizer (Kingma & Ba, 2015) throughout our study. The dropout rate is consistent with the original configuration of the models which is 0.1. For task-specific tuning, we use 8 Nvidia A100 GPUs. For instruction tuning, we use 8 Nvidia V100 GPUs for BASE and LARGE-sized models, 32 for XL, and 64 for XXL. The overall batch size and other detailed hyperparameters for the two settings are in Tab. 5 and Tab. 6, respectively.

 $^{^8}$ https://github.com/facebookresearch/fairseq/tree/main/examples/translation

⁹The signature of sacrebleu for IWSLT14 DE→EN is nrefs:1|case:mixed|eff:no|tok:13a| smooth:exp|version:2.3.1, and for WMT14 EN→DE nrefs:1|case:mixed|eff:no|tok:intl| smooth:exp|version:2.3.1, respectively.

¹⁰https://github.com/CLUEbenchmark/LGEB

[&]quot;https://github.com/google-research/url-nlp/tree/main/mgsm

¹²https://huggingface.co/datasets/Muennighoff/flan

¹³https://huggingface.co/datasets/SirNeural/flan_v2

¹⁴https://huggingface.co/datasets/cais/mmlu

¹⁵https://huggingface.co/datasets/khalidalt/tydiqa-goldp

¹⁶https://huggingface.co/datasets

Table 5: The training hyperparameters for task-specific finetuning.

Dataset	Pretrained model	Batch size (#. tokens)	Learning rate	#. training steps
	XLM-R-BASE	32K	5e-5	150,000
IWSLT14 DE→EN	XLM-R-LARGE	32K	5e-5	150,000
IWSLI 14 DE→EN	XLM-R-XL	32K	5e-5	100,000
	XLM-R-XXL	32K	5e-5	30,000
	XLM-R-BASE	128K	5e-5	300,000
WMT14 Exc. De	XLM-R-LARGE	128K	5e-5	300,000
WMT14 En→DE	XLM-R-XL	128K	5e-5	150,000
	XLM-R-XXL	128K	5e-5	100,000
	XLM-R-BASE	16K	5e-5	30,000
C' 1.10IZ	XLM-R-LARGE	16K	5e-5	10,000
Gigaword-10K	XLM-R-XL	16K	5e-5	5,000
	XLM-R-XXL	16K	5e-5	1,000

Table 6: The training hyperparameters for instruction finetuning.

Training data	Pretrained model	Batch size (#. sequence)	Learning rate	#. training steps
	XLM-R-BASE	512	5e-5	5,000
El 2021	XLM-R-LARGE	512	5e-5	5,000
Flan 2021	XLM-R-XL	512	5e-5	3,000
	XLM-R-XXL	256	5e-5	1,000
	XLM-R-BASE	512	1e-5	70,000
Flan 2022	XLM-R-LARGE	512	1e-5	30,000
rian 2022	XLM-R-XL	1024	1e-5	17,000
	XLM-R-XXL	2048	1e-5	4,000

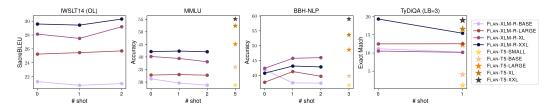


Figure 10: Few-shot performance of Flan-XLM-R and Flan-T5 models. "OL" means the results are obtained with oracle length, while "LB" means the number of length beams to sample the target with length prediction. The model sizes refer to the number of non-embedding parameters.

B FULL EXPERIMENTAL RESULTS

The experimental results for task-specific tuning and instruction tuning on Flan 2022 are in Tab. 7 and Tab. 8, respectively.

Table 7: Full experimental results of task-specific finetuning. OL: the results are obtained with oracle length. LB: the size of length beam for length prediction.

Dataset (Metric)	Setting	XLM-R-BASE	XLM-R-LARGE	XLM-R-XL	XLM-R-XXL
IWSLT14 DE→EN	OL	35.78	38.84	40.11	40.65
(SacreBLEU)	LB=10	34.10	37.33	38.54	38.57
WMT14 EN→DE	OL	26.65	30.22	30.91	32.81
(SacreBLEU)	LB=10	26.72	29.04	30.23	30.34
Gigaword-10K	OL	28.83	31.33	31.72	32.57
(Rouge-L)	LB=10	27.52	30.11	31.42	31.54

Table 8: Full experimental results of instruction tuning on Flan 2022. OL: the results are obtained with oracle length. LB: the size of length beam for length prediction.

Dataset (Metric)	Setting	XLM-R-BASE	XLM-R-LARGE	XLM-R-XL	XLM-R-XXL
	0-shot (OL)	21.26	25.24	28.13	29.59
IWSLT14 DE→EN	2-shot (OL)	20.97	25.70	29.19	30.31
(SacreBLEU)	0-shot (LB=3)	17.76	25.12	26.42	30.90
	2-shot (LB=3)	15.91	23.49	27.29	31.04
MMLU	0-shot	31.28	32.79	40.17	42.13
(Accuracy%)	2-shot	28.74	32.72	38.08	42.06
	0-shot (OL)	23.42	62.28	83.39	84.76
TyDiQA	1-shot (OL)	23.46	66.86	86.12	84.36
(Rouge)	0-shot (LB=3)	13.54	62.28	83.39	84.76
	1-shot (LB=3)	12.63	44.49	55.78	84.36
MGSM (DE)	0-shot	0.9	2.8	1.6	3.6
(Accuracy%)	3-shot	1.6	2.8	5.2	4.4
GSM8K	0-shot	3.6	3.2	5.2	4.4
(Accuracy%)	3-shot	3.2	2.0	3.6	5.6