

Idiomatic Blind Spots in Discrete Diffusion Models

Joshua Chaz Rivera, Vincent Gripon, Bastien Pасdeloup

IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France

{joshua.rivera, vincent.gripon, bastien.pасdeloup}@imt-atlantique.fr

Abstract

Discrete diffusion language models have emerged as a compelling alternative to autoregressive approaches, offering a range of structural advantages. This is due to their bidirectional attention allowing a global semantic view across intermediate denoising steps that enables flexible, order-agnostic generation. Despite this promise, they have received considerably less attention in ongoing research, and existing work has focused narrowly on mathematics and formal reasoning tasks. Their language comprehension capabilities, particularly for figurative language, remain entirely unevaluated, leaving open fundamental questions about how these models handle complex idiomatic structures especially for languages other than English. We find that diffusion models do possess many strong idiomatic processing abilities, demonstrated through competitive token recovery in cloze tests and strong semantic ability from multiple-choice accuracy. However, this advantage fails to surface during inference as confidence-based demasking appears poorly suited for idioms, as early committed tokens override internal representations and discourage certain types of multi-token structures from generation, limiting their linguistic abilities.

1 Introduction

Discrete diffusion language models (dLLMs) have emerged as an increasingly promising alternative to more mainstream autoregressive large language models (Nie et al., 2025). Their iterative denoising process offers distinct advantages over the strict “left-to-right” generation to which autoregressive models are constrained. Specifically, they possess the ability to commit tokens in any order and utilize bidirectional attention throughout the generation process. However, absent from existing prior work is any treatment of figurative language, such as idioms, which we hypothesize would be better supported in diffusion due to their improved handling

of semantics, including in data sparse conditions (Prabhudesai et al., 2025; Zhang et al., 2026). However, the architecture remains relatively immature in comparison, not being widely adopted yet and having fewer available trained examples. The existing literature has overlooked many language tasks in favor of benchmarks focusing on mathematical reasoning or logic tasks (Wang et al., 2026).

In this work, we investigate whether current diffusion models possess the ability to process idiomatic text compared to other LLMs of similar scale, and what advantages or disadvantages this implies for real-world deployment on idiom-related tasks. We examine this through four experiments designed to help disentangle the effects of training data from those of the architecture itself. We examine how readily diffusion models produce idioms in open-ended generation, establishing whether a gap exists relative to autoregressive models. We test whether idiomatic knowledge is present and accessible through two complementary tests, a cloze completion which measures whether idiomatic tokens can be recovered under their full bidirectional context as lexical units, and multiple choice questions which tests whether semantic knowledge is accessible when generation pressure is removed. Finally, probe injection tests whether earlier committed tokens can override internal representations, examining the demasking as a unique mechanistic explanation for deployment failure.

We find that diffusion models show strong idiom retrieval abilities, but fail to generate idioms spontaneously at a comparable rate and are significantly steered by early tokens, especially for cases where idiomatic representation is weaker. These point to a potentially significant limitation in default confidence-based demasking that may make it poorly suited for idiom-related tasks. These considerations currently limit diffusion’s viability as a full replacement for this domain without further improvements to their implementations.

2 Background

Little research has examined diffusion models on figurative language or related domains. As a framework, diffusion has only notably been applied to idiom-related tasks multi-modally through image-generation pipelines paired with autoregressive LLMs, to produce visual metaphors or idiom-based visual puns (Chakrabarty et al., 2023; Xiao et al., 2025), rather than for language modeling in its own right. To our knowledge, diffusion language models have not yet been used in this capacity.

2.1 Diffusion Models

Diffusion models are a class of models that learn to generate data by reversing a gradual noising process (Sohl-Dickstein et al., 2015). During training, a forward process progressively corrupts clean data into noise; a neural network then learns the reverse process, iteratively denoising back to the source. At inference, a model can start from pure or partial noise and recovers a more coherent output through the learned denoising steps. Diffusion has seen strong results in continuous domains such as image generation (Ho et al., 2020; Rombach et al., 2022), motivating its extension to discrete domains such as natural language.

2.2 Discrete and Masked Diffusion Models

Since language is inherently discrete owing to its finite vocabulary and grammatical structure (Ferrone and Zanzotto, 2020), applying diffusion to text requires adapting the framework to a discrete finite token space. Discrete diffusion models (Austin et al., 2021; Hoogeboom et al., 2021) define a forward process over a vocabulary using categorical transition matrices, with a learned reverse model predicting original tokens from corrupted inputs. A particularly effective variant is *masked diffusion* (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024), where the forward process replaces all tokens with a [MASK] token. The reverse model predicts masked tokens from unmasked context, in some ways resembling encoder-based masked language modeling in BERT (Devlin et al., 2019). At inference, the model iteratively demasks tokens from a masked sequence and commits them using a predefined demasking strategy to determine over multiple steps which tokens to continue seeing in the following steps. This strategy is typically based on logit confidence in the demasked tokens at each step, as committing randomly demasked tokens

performs worse on many language-based tasks, despite being the theoretical ideal method (Nie et al., 2025). Alternative denoising strategies are possible, but remain less common and are still largely optimized for mathematical tasks, prioritizing accuracy or speed over improvements in fluency (Tian et al., 2025; Li et al., 2026). Unlike autoregressive models, masked diffusion conditions on arbitrary subsets, enabling bidirectional attention in each step and non-sequential generation.

2.3 Current Availability

Several publicly available discrete diffusion models demonstrate competitive performance against autoregressive alternatives. Full-attention variants have been applied to coding (Gong et al., 2025; Labs et al., 2025) and reasoning tasks (Zhao et al., 2025), while LLaDa (Nie et al., 2025) and Dream (Ye et al., 2025) show that masked diffusion can match the capabilities of autoregressive models even at modest scales. Block diffusion (Arriola et al., 2025) offers a middle ground by dividing the output into semi-autoregressive segments where diffusion is used locally for a fixed range of tokens before moving on to the next block to its right in the output. Building on this, TraDo (Wang et al., 2025) applies a trajectory-aware RL framework to the block-diffusion SDAR architecture, substantially outperforming both LLaDa and autoregressive models like Qwen2.5-7B-Instruct (Team, 2024) on multiple reasoning tasks. Further, LLaDa-2.1 (Bie et al., 2026) introduced a larger 16B parameter model with the ability to do post-commitment token editing, a self-correction capability impossible in autoregressive models by definition. This variety of architectural differences highlights the unique flexibility of diffusion-based approaches.

2.4 Idioms

Idioms are a type of fixed multi-word expression (MWE) that have a figurative, non-literal, reading that are culturally dependent, making them remain a persistent challenge for all types of models (Constant et al., 2017). This difficulty is a consequence of them being non-compositional (Dankers et al., 2022b) which means the meaning of individual words in the idiom does not clearly indicate the meaning of the full string. The degree to which this applies varies across specific idioms (Dankers et al., 2022a). Existing systems often interpret or translate these expressions word-for-word, especially for cases not frequently seen during training (Mi

et al., 2025), with this problem compounded in lower-resource languages. Even higher-resource languages have sparse idiomatic coverage in training corpora since idiomatic expressions make up only a small share of all tokens in the data. Known MWEs are often memorized explicitly by models and retrieved as a whole in early layers for transformer models (Miletić and Schulte im Walde, 2024; Oh et al., 2026), but when not in memory, comprehension correlates with compositionality (Kim et al., 2025). Notably, autoregressive models still fall behind fine-tuned BERT models on figurative language identification tasks (Devlin et al., 2019; Tan and Jiang, 2021; Phelps et al., 2024), a gap attributed to their lack of bidirectional attention and the richer semantics it provides. This makes diffusion a natural candidate for stronger idiomatic processing, yet as of the writing of this paper we have not found any work attempting to evaluate or improve them on this.

2.5 Data

There are no standard datasets, formats, or benchmarks that are used for idiom research and there are substantial differences between the goals for different research domains (Flor et al., 2025). Idiom data is easier to locate for English such as MAGPIE (Haagsma et al., 2020), EPIE (Saxena and Paul, 2020) or FLUTE (Chakrabarty et al., 2022) which all contain high quality English expressions containing idioms, but differ in how they are formatted, generated, and annotated. Significantly less effort has gone into dataset construction for most other languages. Some multilingual datasets are available, such as the Konidioms Corpus (Shaikh et al., 2024), Lidioms (Moussallem et al., 2018), and MIDAS (Kim et al., 2025), which all contain idioms in languages besides English, along with definitions. However, they also contain significant differences in their formats. Lidioms contains all definitions in English for relatively common idiomatic expressions while MIDAS has a much broader definition of what counts as an idiomatic expression with the definitions in the same language as their source, making MIDAS a more challenging dataset. These issues make fair evaluation difficult, but possible.

3 Methods

Models can potentially process idioms accurately in one of two ways depending on the task and the idiom itself. It could be retrieved from its internal

memory of known MWEs or it could attempt some form of intermediate reasoning to infer unknown ones from context. The following experiments test whether there is a generation gap observed in practice and, if there is, whether it reflects a knowledge deficit, a semantic representation failure, or a property of denoising.

We selected multiple publicly available dLLMs for our analysis to be used across multiple tests. Specifically, we used LLaDa (Nie et al., 2025), LLaDa-2.1 (Bie et al., 2026), Dream (Ye et al., 2025), and TraDo’s instruct (I) and thinking (T) variants (Wang et al., 2025) with each of their confidence-based demasking and block or full diffusion specified when relevant. To compare against existing autoregressive models, we also looked at LLaMa-3.1-8B (Grattafiori et al., 2024), Gemma-3-12B (Team et al., 2025), and Qwen2.5-7B-Instruct (Team, 2024) as baselines. For data, we used idioms and definitions in English, Portuguese, Russian and Italian from the Lidioms dataset (Moussallem et al., 2018) in addition to English, Korean, Arabic, and Chinese from the MIDAS dataset (Kim et al., 2025). For each language we used up to 150 idioms per language due to that being the max contained in each Lidioms set. Due to an obvious lack of data and training in truly low resource languages, for this work we focused on evaluation of idioms as a whole including across languages besides English that have weaker presence in the training of these models, but are not strictly low resource by world standards.

3.1 Free Generation

While controlled testing conditions are informative, they would fail to capture the vast majority of real-world scenarios in which idioms or figurative language could arise naturally in generation. Therefore, we first assess some models’ behavior without any explicit request for an idiom in prompting in order to see if there are any notable surface level differences that would motivate further testing.

To achieve this, we prompt each model to produce open-ended narratives in response to a series of 50 questions that have no single correct answer. This means the overall idea, writing style, and narrative direction can potentially vary considerably across model responses. The list of prompts was generated by requesting Claude 4.6 Sonnet to write them and was specified to give them in five genres or registers of request (narrative, opinion, description, advice, and reflective). The variety of prompts

Cat.	Prompt
Narr.	Write about the last day of a beloved local shop before it closes for good.
Op.	Explain why failure is often a better teacher than success.
Desc.	Describe the silence of a forest after a heavy snowfall.
Advice	Explain how someone can rebuild trust after making a serious mistake.
Refl.	Reflect on how grief changes a person over time.

Table 1: Representative sample of free generation prompts, drawn from the prompts used in the experiment. Full prompt list available in the repo.

and genres is intended to reduce bias from a single type of prompt eliciting more figurative language by chance. We test this only in English to avoid issues involving limited training in the other languages. Examples of each are shown in Table 1.

We compare a diffusion model using full diffusion with LLaDa, semi-autoregressive diffusion with TraDo-I using its default block size of four tokens, and an autoregressive baseline with LLaMa, all three of which are 8B parameter models and given a maximum of 200 tokens to answer. We restrict this experiment to one representative model per generation paradigm to serve as an initial comparison and examine paradigm-specific differences rather than model-specific ones, given the variation that exists in the design of diffusion models. The resulting outputs are then passed to Gemma-4-31B and Qwen-3.6-27B, as separate and larger language models, serving as independent judges. The judges are used to scan the output of each model to determine which subtexts are idioms or other figurative language. This is then used to determine an integer count of how many instances occur per generated output. The prompt questions vary by genre and format, allowing a better diversity of outputs to be evaluated, but the same list of prompts is given to all models. The answer overall and other quality concerns were mostly ignored; this was simply a count of the idioms regardless of other factors ¹.

3.2 Cloze Test

Cloze tests (Taylor, 1953), or fill-in-the-blank questions, have been extensively used for tasks evaluating other masked models like BERT (Xie et al., 2018; Cui et al., 2021). They allow us to see how well a model understands a specific concept

by requiring it to complete a missing token in a text (Petroni et al., 2019). With the context provided, top-k token prediction measures whether the model encodes the MWE as a retrievable unit from memory. This allows us to use a comparable experiment setup to (Kim et al., 2025) where we can see whether diffusion models can recover idiomatic tokens more reliably than autoregressive models when given equivalent context visible to the attention, isolating token-level idiom expression retrieval from the full generation.

We examine the model’s output logits for the position corresponding to a target token in the idiom and check whether the correct token appears among the ten highest-ranked predictions. If the correct token exists, we consider the idiom memorized. Autoregressive models, unfortunately, do not have the ability to probe an arbitrary token while allowing both sides to be seen. This makes testing against all possible intermediate generations of diffusion inherently unfair. To provide a comparable situation, we only probe the last token per idiom for both the diffusion models and autoregressive baselines. In both cases, the definition of the idiom was provided on the left side before the idiom as well.

3.3 MCQ

Multiple choice questions are a common way of evaluating a model’s understanding of various concepts (Robinson and Wingate, 2023; Zhang et al., 2024). They allow the models to be simply tested on their understanding by discrimination rather than relying on potentially noisy generations. We set up an MCQ test where we ask multiple models to select one of four possible answers with each answer being a possible definition of the provided idiom. Since we looked at smaller models and cases besides English that might simply lack strong enough representations due to few examples in training, we did not want to confound any results by providing surface level readings or trick answers. We wanted only to see if the model possesses any semantic capacity to discriminate to the correct answer at all. Models were prompted to briefly justify their choice before answering, then responded with a single letter (A–D), from which we extracted the predicted answer via pattern matching; greedy decoding was used for the autoregressive baselines, while diffusion models used their respective default remasking strategies. All models received the same prompt template and had up to 128 tokens

¹Code and data: <https://github.com/JoshuaCRivera/idiom-generation-diffusion>

to respond. The idioms and definitions were taken from Lidioms and MIDAS. Incorrect definitions for possible choices were randomly selected from definitions of other idioms from the same language in the same dataset to ideally provide unambiguously incorrect options, but with the same definition format. Answer choices in the questions were shuffled randomly before testing.

3.4 Directional Probing

Diffusion can see the previous intermediate generation when generating and committing tokens in the following step, causing the final result to be influenced by the earlier and possibly non-sequentially generated tokens. While this is by design, it remains questionable to what degree the retrieval of idiomatic knowledge is governed by its internal representation or if the denoising schedule plays a bigger component. Specifically, if we prompt a model that knows many idioms to varying degrees to explain their meanings, but it outputs a misguided initial start, how influential would this be on its final output or is it able to recover to end with its prior understanding intact? Since this is an idealized situation specifically asking about its meaning, it would also give insight into how the attention mechanism is used in the denoising process, even for more subtle cases and how other tokens in general affect the final result in any generation task. We give the models a prompt with the idiom in full and ask them to explain the meaning of its definition. The prompts clearly indicate that it is its idiomatic usage of the given MWE to avoid ambiguity. We then run a baseline as well as insert two probes to manually replace the masked tokens in the first step and first few tokens of the masked output sequence. To probe, we steer with an inserted string stating that it knows the answer is a chosen definition and let it continue generation from there. The insertion is intentionally not subtle as minor token adjustments might be ignored or insufficient to cause a new output trajectory unless given to an extremely weak model. We use LLaDa-2.1 with its default block size of 32, as it is the largest and strongest diffusion model that we tested and it is less constrained by strict semi-autoregressive generation. This allows us to test whether early-token override persists even where idiomatic knowledge is known to be comparatively strong. We evaluate if the output changed by giving judge model, Gemma-4-31B, the baseline output with one of the two steering probe outputs at a time to compare the

FIGURATIVE EXPRESSION COUNTS			
	TraDo-I	LLaDa	LLaMa
Total	145 / 64	89 / 42	249 / 180
Mean/prompt	2.90 / 1.28	1.78 / 0.84	4.98 / 3.60
Std. dev.	1.85 / 1.89	1.50 / 1.36	3.18 / 3.01
Idioms	1.16 / 0.72	0.62 / 0.48	1.54 / 1.78
Metaphors	1.60 / 0.46	1.12 / 0.28	2.62 / 1.20
Similes	0.04 / 0.04	0.02 / 0.02	0.44 / 0.38
Other	0.10 / 0.06	0.02 / 0.06	0.38 / 0.24

Table 2: Figurative expression counts per prompt in open-ended generation, judged by Gemma-4-31B / Qwen-3.6-27B. Mean and std. dev. are per prompt; subcategory rows show means per prompt.

meaning, and simply state if the output meaningfully changed as a result of the inserted steering. As a baseline, we run the same test with Qwen-2.5-7B, since it reported strong abilities against bigger models and was trained for similar tasks as LLaDa, with the probe inserted after the system prompt to see if the resistance differs.

4 Results

We report results across the four experiments, each focusing on a different stage of idiom processing: whether idioms surface spontaneously in generation, whether their tokens are recoverable from memory, whether their meaning is accessible, and whether that meaning survives the demasking process itself. Together, these results trace a consistent trajectory from strong underlying idiomatic knowledge to its failure to reliably surface in unconstrained generation.

4.1 Free Generation Results

Table 2 reports the figurative language counts produced by TraDo-8B-Instruct (I), LLaDa-8B, and LLaMa-3.1-8B across the open-ended generation prompts, as judged by Gemma-4-31B and Qwen-3.6-27B. The results reveal a substantial gap in the natural use of figurative language between the two model types. Regardless of the judge, LLaMa-3.1-8B produced significantly more figurative expressions than the diffusion outputs. This gap is consistent across all subcategories as well. LLaMa-3.1-8B produces more than twice as many idioms per prompt compared to LLaDa’s full diffusion, and the disparity is even more pronounced for other types of language including similes. The higher variance in LLaMa’s output further suggests that it engages with figurative language more dynami-

TOP-10 RECOVERY — LAST TOKEN					
Source	Dream-v0	LLaDa-8B	TraDo-I	Gemma-3	LLaMa-3
EN-Lidioms	83.3	82.0	84.7	70.7	69.3
RU-Lidioms	58.1	54.3	78.1	76.2	72.4
IT-Lidioms	73.3	58.0	83.3	78.0	81.3
PT-Lidioms	57.4	52.2	79.1	58.3	69.6
EN-MIDAS	78.7	72.0	83.3	44.0	45.3
KO-MIDAS	62.0	76.0	60.0	42.7	46.7
AR-MIDAS	43.3	58.0	61.3	25.3	31.3
ZH-MIDAS	33.3	80.0	84.7	56.7	63.3
Mean	61.2	66.6	76.8	56.5	59.9

Table 3: Top-10 recovery rate (%) for the last token of each idiom across diffusion models and autoregressive baselines (right of vertical rule).

cally, with some prompts eliciting notably richer figurative output while others remain more literal. LLaDa, by contrast, produces a more uniformly sparse and figuratively flat set of responses regardless of the prompt. However, between these two in the results, we see that TraDo, which uses small semi-autoregressive blocks to generate, also shows stronger results compared to LLaDa’s full diffusion, despite similar training objectives and its original evaluation based on mathematical reasoning. TraDo-I’s intermediate position between full diffusion and autoregressive generation is consistent with a gradient in which more autoregressive generation corresponds to more figurative language, though with only three models tested, this pattern should be treated as suggestive rather than an established trend. What is clearer from these results is that both diffusion models produce less dynamic figurative language than the autoregressive baseline despite comparable scale. The results with the two diffusion models raise the question of whether these diffusion models lack idiomatic knowledge due to a less figurative training mix or if they possess it, but fail to deploy it due to something inherent to their design with the denoising process. The lower figurative density of diffusion outputs may also reflect a conservative generation style caused by their distinct generation objective, discouraging idiomatic expressions as they carry lower confidence compared to literal alternatives.

4.2 Cloze Results

Diffusion Comparison Table 3 shows the last token retrieval results across different model for the target token occurring in its top 10 predictions

based on logits. All of the diffusion models clearly show an overall stronger handling of semantic information, where they can better complete known expressions based on prior context. TraDo-I’s advantage in this condition has one plausible explanation, its short semi-autoregressive training objective possibly along with its trajectory-aware fine tuning method. Having the full preceding idiom span visible as context is what that objective normally would reward in math tasks, and the results may suggest the model has learned to exploit idiomatic token sequences as coherent units rather than independent tokens, though we cannot rule out block-size and fine-tuning effects independently. LLaDa and Dream show strong results across the board with both showing comparable results for English as well as some other languages, meaning differences between diffusion models could just be a question of training. Dream-v0’s sharp drop on Chinese (33.3%) and Arabic (43.3%) is harder to attribute to architecture alone since it is an outlier, making it likely reflect sparser coverage of these idioms due to the nature of the MIDAS data’s style.

Diffusion vs. Autoregressive models. Diffusion models all outperform the autoregressive baselines on average, but the margin is not uniform: TraDo-I leads by 17–20 points over LLaMa-3.1 and Gemma-3 and LLaDa is ahead by roughly 7–10 points as well, while Dream-v0 is only marginally stronger. This suggests the advantage is real but model-specific characteristics can cause significant variation. Most likely, diffusion is actively utilizing its distinct training and the attention at its disposal to make better semantic decisions. While diffusion shows more variation, autoregressive models follow a more expected gradient relative to resource level and ease of idiom data. Notably, there are cases where each diffusion model beat Gemma-3-12B despite all of them being smaller models. This is even the case for English, further hinting at an architectural advantage that cannot be attributed to training alone. Since we see overall higher results for diffusion on average, this indicates that its attention is actively utilized leading to better results, when sufficient relevant priors are present. Whether this encoding advantage translates into deployable knowledge during generation is the question the remaining experiments address. Crucially, this advantage is the case when both types of models were given the full context of the idiom and definition, suggesting the gap reflects the distinct

LIDIOMS MCQ							
Source	LLaDa-2.1	Gemma	LLaDa	Dream	TraDo-I	TraDo-T	LLaMa-8B
EN	96.7	98.7	90.7	81.3	94.0	72.0	98.0
RU	63.8	88.6	49.5	53.3	78.1	51.4	69.5
IT	58.7	89.3	53.3	66.7	78.0	54.0	56.7
PT	53.9	71.3	51.3	59.1	64.3	54.8	52.2
Mean	68.3	87.0	61.2	65.1	78.6	58.1	69.1

Table 4: MCQ accuracy (%) on Lidioms. A vertical rule separates the larger models (LLaDa-2.1, Gemma-3-12B) from the 8B-scale models.

training actively enriching token representations from the context provided rather than simply an information asymmetry in what they can access.

4.3 MCQ Results

Tables 4 and 5 report accuracy across all models grouped by scale. Among 8B models, diffusion systems are broadly competitive with LLaMa-3.1-8B and in some cases exceed it. TraDo-I is the clearest example, outperforming LLaMa on both datasets (78.6% vs. 69.1% on average for Lidioms; 85.8% vs. 71.8% on average for MIDAS), including a commanding lead on Arabic and Chinese. Both LLaDa and Dream perform comparably to LLaMa on English as well, but degrade more on the other languages on average, though this is not uniform where Dream beats LLaMa on Portuguese in Idioms (59.1% vs. 52.2%) and both LLaDa and Dream remain within a few points of LLaMa across several other non-English conditions. Crucially, neither type model currently handles these other languages well, presumably due to limited training. Given the variety in results leaves room for improvement for both architectures, this implies limited training in this domain is certainly an important component, regardless of other differences. TraDo-T consistently underperforms TraDo-I despite its added reasoning capacity, suggesting extended chain-of-thought is not beneficial and may be counterproductive for idiomatic MCQ under a math-focused pretraining regime. Overall, we see that diffusion models at the same scale can match or exceed an autoregressive model on this task. With a more controlled training objective, we would presumably see meaningful improvements.

The larger LLaDa-2.1 and Gemma-3-12B expectedly score higher overall, though LLaDa-2.1’s advantage over 8B diffusion counterparts narrows

MIDAS MCQ							
Source	LLaDa-2.1	Gemma	LLaDa	Dream	TraDo-I	TraDo-T	LLaMa-8B
EN	96.0	92.7	81.3	81.3	91.3	66.7	87.3
ZH	96.7	91.3	92.0	82.0	97.3	82.7	80.0
AR	60.7	90.7	43.3	54.0	89.3	61.3	66.7
KO	46.0	77.3	30.7	50.7	65.3	42.0	53.3
Mean	74.9	88.0	61.8	67.0	85.8	63.2	71.8

Table 5: MCQ accuracy (%) on MIDAS. A vertical rule separates the larger models (LLaDa-2.1, Gemma-3-12B) from the 8B-scale models.

on lower-resource languages, confirming that scale alone does not compensate for limited non-English training coverage, although it is a component. All models follow a clear accuracy gradient from high- to lower-resource languages, and where LLaMa holds an average edge over diffusion models it is driven by this coverage gap rather than a fundamental difference in idiomatic reasoning ability. English and Chinese approach ceiling for several diffusion models, with TraDo-I reaching 97.3% on Chinese. Arabic and Korean remain hardest across all model families, with several systems near the random baseline. Chinese is a notable positive outlier for diffusion models across MCQ. This along with the cloze results, lead us to speculate this reflects the partially compositional nature of Chinese characters, which may make their meaning particularly recoverable as discrete units under masked prediction, though differences in tokenization and training data coverage likely also contribute. Together, this confirms that diffusion models retain idiomatic semantic knowledge comparable to autoregressive models at the same scale on average; with neither type possessing perfect results as is, with equally strong English results across the board and only marginal worse results on average across all languages. This shows that semantically they can be improved with a dedicated training objective and is not the primary difference in generation frequency.

4.4 Directional Probe Results

The cloze and MCQ results together establish that diffusion models currently possess idiomatic expressions in memory, but do not spontaneously deploy them as readily in the free generation. The probe injection experiment now asks whether the demasking process itself can actively suppress that

PROBE SUSCEPTIBILITY						
	Baseline		Correct		Incorrect	
	LLaDa	Qwen	LLaDa	Qwen	LLaDa	Qwen
AR	18.0	61.3	60.7	34.0	70.0	48.0
EN-M	70.7	76.7	22.7	18.0	18.0	19.3
EN-L	90.0	94.0	7.3	4.0	18.0	12.7
IT	36.0	58.0	54.0	25.3	64.0	20.7
KO	9.3	22.7	69.3	58.0	60.0	60.7
PT	29.6	46.1	53.9	33.0	65.2	22.6
RU	34.3	67.6	62.9	18.1	75.2	11.4
ZH	75.3	77.3	21.3	20.0	13.3	11.3
All	45.4	63.0	44.0	26.3	48.0	25.8

Table 6: Results for correct final idiom interpretation for LLaDa and Qwen as evaluated by Gemma-4-31B, reported for the no-probe baseline alongside susceptibility rates (%) under correct-probe and incorrect-probe.

knowledge based on the bidirectional attention being too influential. If early committed tokens steer generation away from a model’s internal representation or do not provide the conditions to emerge, this would provide a mechanistic explanation for why idiomatic expressions fail to surface even when the model demonstrably knows them.

Table 6 reports susceptibility rates and prior knowledge across languages for both LLaDa and Qwen, evaluated by Gemma-4-31B. Specifically, we see how influential the probes are with the baseline showing correct definitions printed on a normal run, and the other two columns showing the percentage that changed to match the new incorrect definition after the probe was inserted to steer it. The two models differ substantially in overall susceptibility despite identical evaluation conditions. LLaDa is steered by incorrect probes in 48.0% of cases versus 25.8% for Qwen even though it is a smaller model. The key moderating factor is prior knowledge where both models have strong priors, as in English and Chinese, susceptibility converges and both resist incorrect injection effectively. Where prior knowledge is weaker, the gap widens sharply as Russian (75.2% vs. 11.4%), Italian (64.0% vs. 20.7%), and Portuguese (65.2% vs. 22.6%) show LLaDa dramatically more susceptible despite comparable prior rates, implicating how tokens are viewed in later steps rather than prior knowledge alone.

Table 7 shows this most clearly through idiom profiles. We classify each idiom’s results into profiles based on its behavior across the three conditions. A Blind idiom is one where the model has no prior knowledge and follows the probe regard-

IDIOM PROFILES				
	Blind		Robust	
	LLaDa	Qwen	LLaDa	Qwen
AR	47.3	23.3	7.3	38.0
EN-M	8.0	5.3	63.3	64.0
EN-L	3.3	2.0	76.0	84.0
IT	39.3	8.0	16.7	49.3
KO	46.7	40.0	4.7	13.3
PT	45.2	13.0	18.3	41.7
RU	50.5	3.8	10.5	61.9
ZH	10.0	7.3	72.0	74.0
Mean	30.1	13.2	34.9	53.2

Table 7: Blind and Robust idiom profile rates (%).

less of whether it is correct or incorrect. A Robust idiom is one where the model has prior knowledge and resists the incorrect probe. LLaDa has more than twice the blind-idiom rate of Qwen and substantially fewer robust idioms, meaning a larger share of its representations lack the stability to resist a misguided early commitment. Qwen’s higher robust rate reflects a model that retains its prior under adversarial pressure; LLaDa more readily treats early tokens as authoritative and generates to rationalize them. Korean is the exception as both models show high incorrect-probe susceptibility and blind-profile rates with nearly no architectural gap, suggesting vulnerability here is driven entirely by data sparsity rather than decoding strategy.

The overall pattern points to confidence-based demasking as the core faulty mechanism. At each demasking step, the model scores all remaining masked tokens and commits those with highest confidence first. For idioms, this creates a structural problem as their weaker, rarer knowledge representations do not resist well against alternative provided information, while something already agreeing with the existing output will be easier and safer for the model to commit. Once those early tokens are committed, bidirectional attention conditions all subsequent masked positions on them, making the idiomatic reading increasingly inconsistent with what is fixed. When we inject a correct or incorrect definition early, we artificially supply the kind of strong early signal that idiomatic tokens normally fail to generate on their own. The high susceptibility of LLaDa relative to Qwen then reflects not just weaker priors but a decoding process with no mechanism to revisit an early commitment once surrounding tokens have rationalized it, unlike autoregressive generation, where a strong prior can seemingly better surface without competing against

the context provided already. Stronger training can clearly compensate, but the improved semantics provided by the bidirectional attention can inadvertently create overly cautious and agreeable models mid-generation that potentially negate its utility.

This may also explain why the diffusion models generated fewer idioms in free generation. Even when there is no semantic priming for idioms or their definition, the nature of confidence based demasking forces the text in a direction that does not provide the conditions for them to arise as easily. Literal alternatives would be valid options that easily agree with the text without the uncertainty that a multi-word idiom, whose literal word senses are irrelevant, would provide. In practice, this blocks the semantics of idioms when not overtly relevant and if there is uncertainty they will be steered in the wrong direction to justify themselves as a less confident component of their generation.

5 Conclusion

We presented an evaluation of discrete diffusion language models on figurative language tasks spanning multiple languages and resource levels, comparing them against autoregressive baselines across four experimental conditions: natural generation, cloze recovery, multiple choice, and directional probing. Our results paint a consistent picture that diffusion models lag behind autoregressive models in idiomatic usage, but possess genuine and measurable structural strengths that their current objectives fail to exploit. We first show that full diffusion models produce substantially less figurative language in free generation than autoregressive models or even semi-autoregressive diffusion approaches, a gap that persists regardless of whether low-confidence or random demasking is used. The cloze results then demonstrated that this is not a knowledge gap as diffusion models have superior token recovery due to their bidirectional semantics, outperforming autoregressive baselines even when both are given the full left context of the idiom. The MCQ results confirmed that diffusion models retain meaningful semantic knowledge and can discriminate against incorrect definitions at rates comparable to autoregressive models for well-represented languages. Finally, the probe injection results revealed that confidence-based demasking cascades early committed tokens in a direction that overrides internal representations, particularly for languages and idioms where prior knowledge is weaker and

a literal early commitment is more likely to go uncorrected.

Critically, confidence-based demasking is not the mathematical ideal for diffusion, but real world linguistics constraints prefer it as a method over random demasking due to fluency and coherence reasoning. This means that this strategy is already an imperfect solution; however it can be improved with some method that better discourages generation based on tracking tokens attending to each other without over-reliance on single-token confidence. Autoregressive models avoid this problem by not having to determine which position to generate at each step.

While diffusion currently has the potential to be competitive against autoregressive models, especially with more training, their current framework and the demasking strategies they are trained around are inherently poorly matched for idiomatic generation. It remains to be tested whether we can improve their ability here without degradation of other abilities and using a training-free approach.

6 Limitations

Diffusion and autoregressive models differ substantially in architecture, making direct comparison difficult without artificial constraints. The cloze test has no perfect way to fairly represent the intermediate generation of both model types given their inherently different generation methods, and the probe injections similarly required proxies reflecting architectural differences with no direct equivalents. The variations we chose were viewed as adequate approximations that would both surface meaningful differences and fairly elicit the effects we wanted to compare.

Each experiment evaluates idiomatic knowledge in a different context by design, which carries its own interpretive limitations. Free generation requires spontaneous idiom deployment with no definition provided, while the cloze, MCQ, and probe tasks all present idioms with explicit definitions in controlled settings. We use the controlled tasks to isolate specific components (token recovery, semantic discrimination, and demasking susceptibility) as potential explanations for the generation gap, but free generation is influenced by additional factors these tasks do not capture such as the overall training used, which probably also pushed the models towards more literal outputs in line with expected requirements for other reasoning tasks.

Training our own models from scratch could alleviate this issue, but would be costly. That said, the convergence of results across cloze and MCQ in multiple languages and resource levels suggests that familiarity with idioms is not the primary bottleneck, especially for English, pointing toward the confidence-based demasking process as the more likely explanation. While the probe’s explicit context differs substantially from natural generation, its strong effect on the controlled condition suggests it is inevitably a factor in less overt settings like free generation too.

Within the diffusion models tested, variation in training objectives and denoising strategies makes comparison difficult. This is why we kept models at their default recommended parameters unless otherwise stated. It is also why we only used one diffusion model for the probing task and did not use all the models for each task. It also explains why we did not use LLaDa-2.1’s remasking ability, which has no analogy in any other model. Future work could leverage this, though it adds complexity and would not alone encourage idiom generation without more fundamental reframing as discussed above. Using LLMs as a judge also adds uncertainty in their results, but we had no alternative method of evaluation of some of these tests. Finally, while we identify limitations of the confidence-based demasking methods and believe improvements via more strategic use of logits are possible, we do not propose a new method at this time.

A Any Position Cloze

Table 8 shows the results if we were to check MWE retrieval from the cloze test, but average across all tokens within the idiom masked one at a time. While not able to directly contrast with autoregressive models, we see similar or better results than the strict end of idiom cloze used to compare. This further shows the architectural advantage that an idiom can be generated out of order and be retrieved in diffusion at a better rate, but requires the idiom to have started to generate which we cannot currently control in normal generation without steering or prompting.

B LLM Judge Inter-annotator Agreement

The LLMs as a judge for the free generation task’s statistics can be seen in Table 9. The results show that the judges have a high correlation between

TOP-10 RECOVERY– ALL TOKENS					
Source	LLaDa	Dream	TraDo-1	LLaMa-3	Gemma-3
EN-Lidioms	89.3	91.0	83.8	69.3	70.7
RU-Lidioms	63.8	63.7	69.4	72.4	76.2
IT-Lidioms	65.5	74.0	71.8	81.3	78.0
PT-Lidioms	52.4	56.5	62.9	69.6	58.3
EN-MIDAS	78.3	84.1	75.3	45.3	44.0
KO-MIDAS	68.4	53.8	46.4	46.7	42.7
AR-MIDAS	56.8	40.9	42.3	31.3	25.3
ZH-MIDAS	78.2	35.1	75.6	63.3	56.7
Mean	69.1	62.4	65.9	59.9	56.5

Table 8: Top-10 recovery rate (%) across diffusion models and autoregressive baselines (right of vertical rule). Diffusion model results are from the any-position masking condition; autoregressive baseline results are only from the last-token condition.

Metric	Score
Gemma Judge Mean Count	3.22
Qwen Judge Mean Count	1.91
Mean Difference (J1–J2)	+1.31
Exact Match	0.280
MAE	1.500
Pearson r	0.819
Spearman ρ	0.724
Cohen κ (binned)	0.377
Zero/Non-zero Agreement	0.700
Paired t -statistic	10.29
p -value	<0.0001

Table 9: Inter-annotator agreement on figurative language count between Judge 1 and Judge 2 (Qwen) across 150 matched texts. A significant systematic bias is observed, with Gemma consistently assigning higher counts than Qwen.

each other, but differ further on the Cohen κ and the percentage of exact matches. This means that while each judge independently agrees on the ratio, they differ more substantially on how many should count. In practice, we see Gemma classifies more substrings as figurative or idiomatic than Qwen. We find that both agree that LLaMa’s generation is more figurative, but disagree as to what extent. Likewise, generally, the idioms that Qwen locates seem to be a subset of Gemma’s identifications rather than a completely separate list. What counted as idioms was not overtly specified in the prompt by us, this means the models had a high degree of discretion to determine what is valid as an idiom.

IDIOM PROFILES				
	Partial		Resistant	
	LLaDa	Qwen	LLaDa	Qwen
AR	13.3	10.7	22.7	24.7
EN-M	14.7	12.7	10.0	14.0
EN-L	4.0	2.0	14.7	10.7
IT	14.7	17.3	24.7	12.7
KO	22.7	18.0	13.3	20.7
PT	8.7	20.0	20.0	9.6
RU	12.4	14.3	24.8	7.6
ZH	11.3	12.7	3.3	4.0
All	12.9	13.2	16.3	13.3

Table 10: Partial and Resistant idiom profile rates (%) for LLaDa and Qwen, completing the profile breakdown from Table 7. **Partial**: susceptible to correct probe only. **Resistant**: prior knowledge present but susceptible to incorrect probe. The four profiles together account for 94.2% (LLaDa) and 92.9% (Qwen) of idioms; the remainder (~6–7%) showed no prior knowledge and adopted neither probe, and are excluded.

C Partial Profile Distributions

Table 10 reports the complete idiom profile breakdown including Partial and Resistant categories omitted from the main text. Both profiles are nearly identical across LLaDa and Qwen (~13% each), consistent with these categories reflecting idiom-level properties in training data rather than architectural differences. Taken together with Table 7, the four profiles account for 94.2% of LLaDa idioms and 92.9% of Qwen idioms; a residual ~6–7% showed no prior knowledge and resisted both probes, suggesting a small subset of idioms whose representations are insufficiently formed to respond to any injected context.

Acknowledgments

This work is funded by the « Agence Nationale de la Recherche » through the France 2030 program under the reference number ANR-23-IACL-0009.

References

- Marianne Arriola, Subham Sekhar Sahoo, Aaron Gokaslan, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Justin T Chiu, and Volodymyr Kuleshov. 2025. [Block diffusion: Interpolating between autoregressive and diffusion language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. [Structured denoising diffusion models in discrete state-spaces](#).

In *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993. Curran Associates, Inc.

- Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fuyuan Chen, Kun Chen, Lun Du, Daozhao Feng, Haibo Feng, Mingliang Gong, Zhuocheng Gong, Yanmei Gu, Jian Guan, Kaiyuan Guan, Hongliang He, Zenan Huang, Juyong Jiang, Zhonghui Jiang, Zhenzhong Lan, and 31 others. 2026. [Llada2.1: Speeding up text diffusion via token editing](#). *Preprint*, arXiv:2602.08676.

- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. [FLUTE: Figurative language understanding through textual explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. [I spy a metaphor: Large language models and diffusion models co-create visual metaphors](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Multiword expression processing: A survey](#). *Computational Linguistics*, 43(4):837–892.

- Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2021. [On commonsense cues in BERT for solving commonsense tasks](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 683–693, Online. Association for Computational Linguistics.

- Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022a. [The paradox of the compositionality of natural language: A neural machine translation case study](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. [Can transformer be too compositional? analysing idiom processing in neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL-HLT*.

- Lorenzo Ferrone and Fabio Massimo Zanzotto. 2020. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI*, 6.
- Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 90–100, Hannover, Germany. HsH Applied Academics.
- Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *The Thirteenth International Conference on Learning Representations*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.
- Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.
- Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*.
- Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.
- Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, and Volodymyr Kuleshov. 2025. Mercury: Ultra-fast language models based on diffusion. Preprint, arXiv:2506.17298.
- Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. 2026. Diffusion language models know the answer before decoding. Preprint, arXiv:2508.19982.
- Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.
- Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.
- Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Jirong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2026. Tug-of-war between idioms’ figurative and literal interpretations in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2942–2958, Rabat, Morocco. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
- Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Kateřina Fragkiadaki, and Deepak Pathak. 2025. Diffusion beats autoregressive in data-constrained settings. Preprint, arXiv:2507.15857.
- Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question

- answering. In *The Eleventh International Conference on Learning Representations*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). *Preprint*, arXiv:2112.10752.
- Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. 2024. [Simple and effective masked diffusion language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue*, pages 87–94, Cham. Springer International Publishing.
- Naziya Mahamdul Shaikh, Jyoti D. Pawar, and Mubarak Banu Sayed. 2024. [Konidioms corpus: A dataset of idioms in Konkani language](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9932–9940, Torino, Italia. ELRA and ICCL.
- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. [Simplified and generalized masked diffusion for discrete data](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. [Deep unsupervised learning using nonequilibrium thermodynamics](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.
- Minghuan Tan and Jing Jiang. 2021. [Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.
- Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Runchu Tian, Junxia Cui, Xueqiang Xu, Feng Yao, and Jingbo Shang. 2025. [Finish first, perfect later: Test-time token-level cross-validation for diffusion large language models](#). *Preprint*, arXiv:2510.05090.
- Guanghan Wang, Gilad Turok, Yair Schiff, Marianne Arriola, and Volodymyr Kuleshov. 2026. [d2: Improved techniques for training reasoning diffusion language models](#). *Preprint*, arXiv:2509.21474.
- Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. 2025. [Revolutionizing reinforcement learning framework for diffusion large language models](#). *Preprint*, arXiv:2509.06949.
- Kelaiti Xiao, Liang Yang, Dongyu Zhang, Paerhati Tulajiang, and Hongfei Lin. 2025. [Visual puns from idioms: An iterative llm-t2im-mlm framework](#). *Preprint*, arXiv:2511.22943.
- Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. [Large-scale cloze test dataset created by teachers](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. [Dream 7b: Diffusion large language models](#). *Preprint*, arXiv:2508.15487.
- Zeyang Zhang, Chengwei Liang, Xingyan Chen, Meiqi Gu, Minrui Luo, Jingzhao Zhang, and Tianxing He. 2026. [Differences in text generated by diffusion and autoregressive language models](#). *Preprint*, arXiv:2605.12522.
- Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. 2024. [Multiple-choice questions are efficient and robust llm evaluators](#). *Preprint*, arXiv:2405.11966.
- Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2025. [d1: Scaling reasoning in diffusion large language models via reinforcement learning](#). *Preprint*, arXiv:2504.12216.