SCALING AND DISTILLING TEXT EMBEDDINGS FOR BETTER DIFFUSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion language models (DLMs) offer a promising alternative to autoregressive (AR) language generation. Recent advances in continuous DLMs, which apply latent diffusion to continuous text embeddings, raise a practical question: which embedding makes the best latent space, i.e., the most diffusible? To answer this, we search through different embeddings and find that *scaling* up to the advanced and general-purpose T5Gemma-2 embeddings greatly improves generative performance. But such raw embeddings are not optimal, as they are overly discriminative and even the embeddings of plausible alternative words are separated. Consequently, continuous diffusion often fails to reach any of them and ends up at an invalid embedding in between. To address this, we *distill* T5Gemma-2 into a student encoder that learns the teacher's decoded probabilities as soft labels. Learning from such soft labels makes the student pull the alternative embeddings closer while maintaining the encoding-decoding mechanism. The distilled embeddings form a more connected and diffusible latent space, improving over the vanilla T5Gemma-2 embeddings. As a result, our medium-sized DLM achieves **Gen. PPL 17.8** (against real-text PPL 15.4) at real-text entropy on OpenWebText, outperforming GPT-2-M as an AR counterpart.

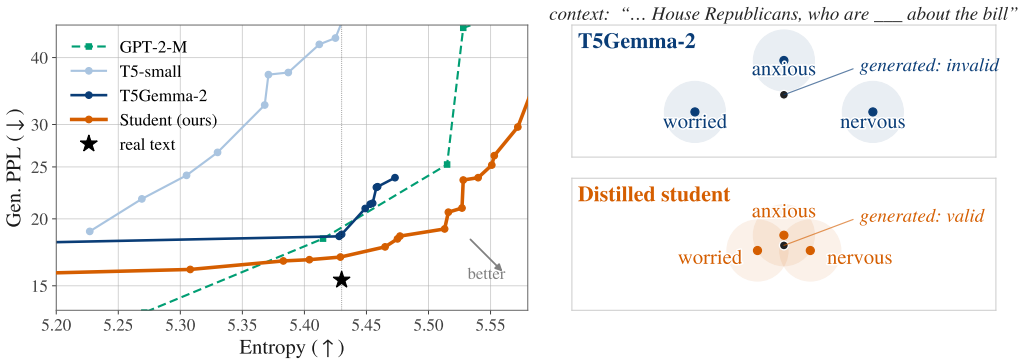


Figure 1: **Overview of our performance and mechanism.** *Left*: the same DLM trained on our embeddings (distilled from T5Gemma-2) reaches real-text entropy with the best quality. *Right*: T5Gemma-2 maps plausible candidates apart in the embedding space, so a generated embedding often lands in the gap between them. Our distilled embedding pulls the candidates closer, forming a continuous region where an embedding at least decodes to one of them.

1 INTRODUCTION

Diffusion models (Ho et al., 2020) have achieved huge success in continuous modalities such as image (Labs, 2025) and video (Seed, 2026). A natural question is whether their strong modeling transfers to language generation, and whether Diffusion Language Models (DLMs) can serve as a more parallelizable and efficient paradigm beyond autoregressive (AR) generation. There have been

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

two families of DLMs: the first is discrete DLMs (Austin et al., 2021; Nie et al., 2025), which apply a discretized diffusion process directly on tokens, and use likelihood training (Lou et al., 2023). They have drawn the most attention since they can be adapted from foundation models and scale with them (Ye et al., 2025; Team et al., 2026). The second is continuous DLMs (Li et al., 2022) which we study in this paper. They apply latent diffusion on text embeddings and decode generated embeddings back to tokens, and are trained with MSE loss. Interestingly, though introduced as early as discrete DLMs, continuous DLMs remain less studied until recently (Hu et al., 2026; Guo et al., 2026), along with techniques including self-conditioning (Chen et al., 2022), diffusion transformers (Peebles & Xie, 2023; Li & He, 2026), and the adoption of pretrained embeddings (Meshchaninov et al., 2026a).

A central question for continuous DLMs is which text embedding to use as the latent space. The same question is already a major topic in image diffusion (Yao et al., 2025; Black Forest Labs, 2025), where the recent Representation Autoencoder (RAE) (Zheng et al., 2026) finds that replacing the VAE latents (Rombach et al., 2022) with pretrained image embeddings improves diffusibility. For text, earlier work finds that pretrained embeddings improve continuous DLMs over simple embeddings from lookup tables (Lovelace et al., 2023; Hu et al., 2026). However, existing approaches either modify the embeddings and the diffusion model jointly (Meshchaninov et al., 2026b; Li et al., 2026) or rely on relatively older embedding models (Lemercier et al., 2026), making it difficult to isolate the contribution of the embedding space or fully realize its potential.

In this paper, we study the embedding space for continuous DLMs in isolation. We fix the diffusion side as the recent Embedded Language Flow (ELF) (Hu et al., 2026) framework for its minimalist design, and change only the embedding. We find two things. **(i) Scaling the embeddings leads to a more diffusible latent space**. In smaller-scale experiments, we find a scaling behavior in the T5 models where diffusibility increases along the lineage of T5-small/base (Raffel et al., 2020) (used in ELF) $\rightarrow$ T5Gemma-1 $\rightarrow$ T5Gemma-2 (Zhang et al., 2025b), and replacing the default T5-small embedding in ELF with T5Gemma-2-270M embeddings reduces generative perplexity by about 40% at the same entropy. **(ii) Scaled embeddings can be further enhanced via distillation.** A student encoder trained to match the probabilities from the T5Gemma-2 decoder is more diffusible than the teacher. As a result, the same DLMs trained on the distilled embeddings improve over the teacher on perplexity and MAUVE and sample more stably, as illustrated in Fig. 1.

The reason that distillation improves diffusibility is that raw T5Gemma-2 embeddings are hard to generate perfectly. When sampling is insufficient or imperfect, the trajectories end at invalid embeddings that deviate from the real ones and decode to an uncertain or even wrong word (Shabalin et al., 2026). We relate this to the strong representation learning of T5Gemma-2, which makes its embeddings information-dense and distinct, so even the words that could fill the same position are distinct (Yu et al., 2020). That helps discriminative tasks but hurts generation, since a continuous trajectory has to settle on one of several separate candidates and can end in between. The soft labels in distillation mitigate such discrepancy, since the student learns to predict the plausible candidates so the alternative embeddings are placed near each other, and the generated embedding lies in the region they form.

Our results also shed light on understanding language representation learning and generation. Although text embeddings are considered contextualized and continuous, they inherit discreteness from text and do not fill the entire embedding space, so continuous DLMs can still generate embeddings that do not represent any word. This, in turn, is the other side of continuous language generation. Though more flexible than predicting tokens, they also risk deviating and generating invalid embeddings. Whether continuous DLMs will become a fully workable route still needs to be tested on even larger and practical scales, and the scaling and distillation here provide a basis for that. In all, our contribution can be summarized as follows:

- We show that scaled embeddings make good latent spaces for continuous DLMs, providing a strong baseline and starting point for continuous DLMs with T5Gemma-2 embeddings.
- We study the diffusibility of text embeddings from the perspective of generating valid, decodable embeddings.
- We enhance T5Gemma-2's diffusibility by distilling its decoder probabilities, providing insights for taming and incorporating foundation models for continuous DLMs.

2 PRELIMINARY AND RELATED WORK

2.1 PRELIMINARIES

Continuous DLMs are latent diffusion on text embeddings. We follow the ELF (Hu et al., 2026) setup and first map the input token sequences $\boldsymbol{s} \in \mathcal{V}^L$ to the sequences of embeddings with a (frozen) embedding model:

$$\boldsymbol{x} = \text{Emb}(\boldsymbol{s}) \in \mathbb{R}^{L \times d}$$

This is also the only part we change throughout the paper.[1] Training and sampling are standard diffusion, where we train a denoiser $\boldsymbol{x}_\theta(\boldsymbol{z}_t, t)$ on embedding sequences corrupted by sequence-wise Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^{L \times d}$,

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{\epsilon}}\Big[\frac{1}{(1-t)^2}\|\boldsymbol{x}_\theta(\boldsymbol{z}_t, t) - \boldsymbol{x}\|^2\Big], \quad \boldsymbol{z}_t := t\boldsymbol{x} + (1-t)\boldsymbol{\epsilon},$$

and the optimal denoiser is the posterior mean:

$$\boldsymbol{x}^\star(\boldsymbol{z}_t, t) = \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{z}_t] = \sum_i p(\boldsymbol{x}_i \mid \boldsymbol{z}_t)\, \boldsymbol{x}_i, \tag{1}$$

where $\{\boldsymbol{x}_i\}$ are the embeddings of the training data, and weights $p(\boldsymbol{x}_i \mid \boldsymbol{z}_t)$ decrease with $\|\boldsymbol{z}_t - t\,\boldsymbol{x}_i\|$.

We also train a decoder that maps the generated embeddings back to tokens. It shares weights with the denoiser $\boldsymbol{x}_\theta$ plus a trainable head, and is trained on embedding sequences with token-wise corruption, where $\boldsymbol{\lambda} \in [0,1]^L$ is a per-token noise level

$$\mathcal{L}_{\text{CE}} = \text{CE}(\text{Head}_{\theta'}(\boldsymbol{x}_\theta(\tilde{\boldsymbol{z}}, t=1)), \boldsymbol{s}), \quad \tilde{\boldsymbol{z}} = \boldsymbol{\lambda}\,\boldsymbol{x} + (\mathbf{1} - \boldsymbol{\lambda})\,\boldsymbol{\epsilon} \tag{2}$$

This improves robustness for decoding imperfect embeddings. We also keep specific design choices of ELF (noise scheduling, self-conditioning, SDE sampling, etc.) unchanged; see App. B.1.

2.2 RELATED WORK

Continuous DLMs and their latents. Earliest continuous DLMs adopt simple and co-trained embeddings (Li et al., 2022). However, co-training collapses the embeddings, since one point for all is the easiest way to reduce the MSE loss (Dieleman et al., 2022; Gao et al., 2024). Recent work finds that pretrained text embeddings perform better (Meshchaninov et al., 2026a; Hu et al., 2026); this paper follows this line and tests their full potential.

Discrete DLMs. Discrete DLMs use a generalized diffusion process with categorical corruption on the tokens (Lou et al., 2023; Ou et al., 2024; Sahoo et al., 2024). Importantly, they can be adapted from AR foundation models and benefit from their progress (Gong et al., 2025; Team et al., 2026). We show that continuous DLMs can co-evolve with them too, through embedding models that are adapted from such foundation models. There are also papers showing discrete DLMs can be enhanced via continuous text embeddings (Zhou et al., 2025; Lemercier et al., 2026).

Diffusibility of embeddings. Which latent is easy to diffuse is a central question in image diffusion (Skorokhodov et al., 2025; Xu et al., 2026), and replacing VAE latents with pretrained representations such as DINOv2 (Oquab et al., 2023) improves diffusibility (Zheng et al., 2026; Shi et al., 2026). We find the text counterpart, where the scaled T5Gemma-2 provides a strong latent space, and instead of explicitly redesigning (Zhang et al., 2023; Jiang et al., 2026) the latent, we distill the scaled encoder for diffusibility.

[1] Strictly speaking, we also change the final LM head since different embeddings have different vocab sizes.

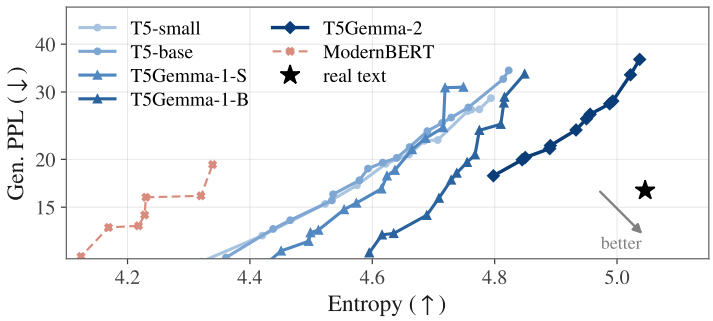


Figure 2: **Performance of the same ELF-B trained with different encoders on OWT-512.** T5Gemma-2 reaches the entropy of real text, and its Gen. PPL is about 40% lower than T5-small's at the same entropy. Within the T5 family, T5Gemma-2 $>$ T5Gemma-1 $>$ T5-small.

3 Scaling Text Embedding for Continuous DLMs

(1) Scaled embeddings are high-potential latents. Embeddings trained with richer data and a better recipe can expose more information for the diffusion to learn, and scaling from T5-small to T5Gemma-2 cuts Gen. PPL by about 40% at matched entropy.
(2) Scaled embeddings can still be hard to generate. Generated embeddings can be invalid and far from real ones. This is because T5Gemma-2 keeps the plausible words for a position apart, so a generated embedding lands between them and decodes problematically.

3.1 Scaling improves the diffusibility of embeddings

We start with off-the-shelf embedding models, on which we train the same ELF-B on a 25% subset of OpenWebText (Gokaslan et al., 2019) (512 tokens per sequence, denoted as 'OWT-512') for faster comparison. We mainly study the family of encoder-decoder (Sutskever et al., 2014) models, including T5-small (Raffel et al., 2020) used in ELF, T5-base, T5Gemma-1 (Zhang et al., 2025a), T5Gemma-2-270M (Zhang et al., 2025b) (denoted as T5Gemma-2 throughout the paper); and we also include a strong modern encoder-only model, ModernBERT (Warner et al., 2025). The embeddings are normalized by their global mean and variance, as in ELF.

We find that as the embedding models scale from T5-small through T5Gemma-1 to T5Gemma-2, their diffusion performance (shown in the PPL-entropy curves in Fig. 2) also improves. T5Gemma-1-B can outperform T5-small, while T5Gemma-2 further improves over T5Gemma-1-B. This suggests that stronger representation learning exposes richer information from the training data for the diffusion to learn. Notably, while the encoder-only ModernBERT also performs well on downstream tasks, it is less suitable for diffusion and generates repetitive sentences.

Here, T5Gemma-2 is the only embedding space able to generate sentences at real-text entropy and cut Gen. PPL by about 40% at the same entropy as T5-small. We therefore pick T5Gemma-2 as the baseline for the rest of the paper. Here we focus on smaller-sized embedding models for a controllable experiment size, while bulky encoders (T5Gemma-2-1B/4B) are left for future work.

3.2 Scaled embeddings are hard to generate perfectly

However, when we inspect the generated embeddings, we find a failure mode: a generated embedding can be far from every candidate word. We take the candidates of a position from the deployed decoding head in equation 2, since its top-k are the words it would decode to.[2] Such an embedding is still decoded, but often with low confidence or to a wrong word: the decoder's top-1 falls with the distance.

[2]The pretrained T5Gemma-2 decoder picks similar words; see App. A.2.

We call a generated embedding *invalid* when no candidate lies within reach. Formally, with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ the embeddings of the top-k candidates, $\hat{\boldsymbol{x}}$ is invalid iff $\min_i \|\hat{\boldsymbol{x}} - \boldsymbol{x}_i\| > 0.91\,\mathrm{nn}$; we also flag positions with top-1 below 0.6 as *uncertain*. Both thresholds are derived from real text references, where the top-1 decoded probabilities from real embeddings are above 0.98, and the nearest neighbor distance is 1 nn by definition. So an embedding within 0.91 nn of a candidate is closer to that word than real embeddings are to each other, and a top-1 below 0.6 means the decoder does not produce a clear majority word. The exact threshold is not critical, as long as it divides the teacher's bimodal distances (Fig. 3). We measure over 64 generated sequences and collect the distance from the generated embedding to its nearest candidate and the decoder's top-1 probability for each uncertain position, and Fig. 4 further shows a real sequence in 2D. We sample at NFE 32, below the sampling grid in Sec. 5, to expose the failure mode; and we see it is not an artifact of the low NFE, as the gap persists across NFE 64–512 (Fig. 8).

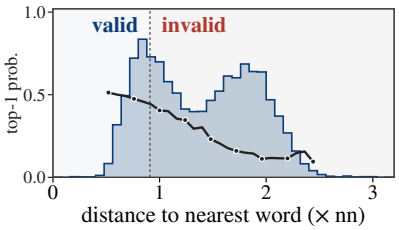


Figure 3: **Generated embeddings deviate from real T5Gemma-2's.**

T5Gemma-2, decoded (NFE 32) PPL 35.9, H 5.50 (real 15.4 @ 5.43)

It may come as a small surprise ing if he doesn ' t agree with those kinds of divergent values , but if his party fails to do what it said it would , the progressive movement could swing in the other direction . WASHINGTON , Vt . -- Former Jersey State Senate member Robert Bridgewater pleaded guilty Wednesday to charges of illegal banking and fraud , after he was investigated in an alleged plan to pro raise Republican candidates . · U . S . Eastern District Court Judge Jay Schroeder gave Brennan three years in prison upon plea . The fines will range anywhere from $ 2 5 , 0 0 0 to $ 5 0 0 , 0 0 0 .

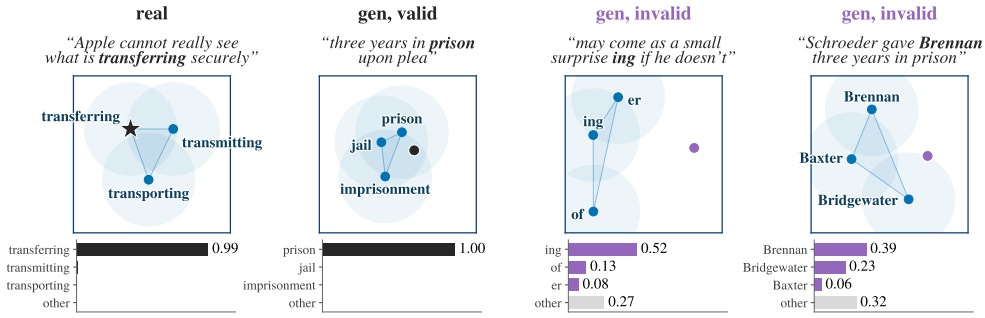


Figure 4: **Visualizing the embedding error in 2D.** *Top:* a sample generated at NFE 32; uncertain tokens (top-1 below 0.6) in purple coincide with the grammar mistakes. *Bottom:* four positions in the plane of their candidate words and the decoded probabilities. The disk marks the 0.91 nn range of a candidate, and an invalid one lands outside every disk and is decoded problematically. See details in App. B.4.

Previous works have studied such embedding errors as interpolation effects (Chen, 2026; Ashiq et al., 2026) or mode-averaging (Aithal et al., 2024), and a common finding is that discrete and separated targets are hard for continuous diffusion to generate (Shabalin et al., 2026). We reiterate here that (ideally) sampling is guided by the optimal denoiser Equation 1, which is a weighted mean of candidate embeddings. If candidates are separated, the mean falls between them and provides a vague direction, and an imperfect sampling trajectory can end at an invalid location far from any of them. The decoder is then forced to predict with such embeddings, causing errors in the final generation.

4 Enhancing the Diffusibility via Distillation

(3) Distilling with soft labels makes the embeddings easier to generate. A student that learns the teacher's decoded probabilities places the plausible words closer to each other; the diffusion then lands on concrete words more often and generates better.

4.1 DISTILLING T5GEMMA-2'S DECODED PROBABILITIES

Why are T5Gemma-2 embeddings hard to generate? We relate it to how T5Gemma-2 is trained. As an encoder-decoder model, it mostly predicts its input as one-hot labels. This maximizes the likelihood of the input token against every alternative, pushing its embeddings away from semantically similar candidates that could fill its place (Wu & Papyan, 2024). This helps with learning rich and distinctive representations (Yu et al., 2020), but hinders similar embeddings from forming a region and leads to invalid embeddings, as a trade-off between discrimination and generation.

Therefore, we distill T5Gemma-2 into a student encoder that learns the teacher decoder's probabilities as soft labels (Fig. 5). The soft labels spread over the plausible candidates of a position, so the student places these candidates closer (Hinton et al., 2015; Müller et al., 2019). For a training sequence $\boldsymbol{s} \in \mathcal{V}^L$, we minimize

$$\mathcal{L}_{\mathrm{KD}}(\theta) = \sum_{j=1}^{L} \mathrm{KL}\big(p_j^{\mathrm{T}} \,\|\, p_j^{\mathrm{S}}\big), \quad p^{\mathrm{T}} = \mathrm{Dec}^{\mathrm{T}}(\mathrm{Enc}^{\mathrm{T}}(\boldsymbol{s})), \quad p^{\mathrm{S}} = \mathrm{Dec}^{\mathrm{T}}(\mathrm{Enc}^{\mathrm{S}}_{\theta}(\boldsymbol{s}))$$

where p_j^{T} and p_j^{S} are the frozen teacher decoder's distributions at position j. We distill on the full OWT, initialize the 9-layer student with every second of the teacher's 18 layers (Sanh et al., 2019), and keep the top-256 candidates of the 262k vocabulary with renormalized probabilities; see App. B.2.

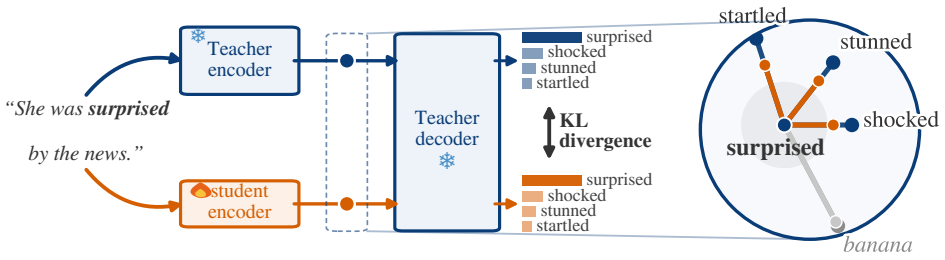


Figure 5: **Illustration of the distillation pipeline and the effect on the embeddings.** *Left:* The student encoder is trained to match the teacher's decoded probabilities (soft labels). *Right:* The student pulls the embeddings of plausible candidates closer while keeping the irrelevant word far.

Ablations For the objective, replacing the KL on soft labels with CE on the one-hot input or MSE on the teacher's embeddings leads to students that never reach real-text entropy (CE$\leq$ 4.18, MSE$\leq$ 5.26, see Fig. 10 in Appendix). For the layer choice, we find that students initialized from the full 18 layers to the 9 we use consistently match or outperform the teacher, so the improvement is not due to reduced capacity. A randomly initialized student does not converge. See App. A.3 for details.

Though we only distill on OWT, the student does not collapse to it. It is still able to encode and decode other text sources, e.g., WikiText (Merity et al., 2016), as shown in App. A.4. As expected, the distillation costs some discriminative power and degrades downstream classification performance (89.4 to 78.1 on SST-2 (Wang et al., 2018)), but it increases the diffusibility as we show in the following section.

4.2 DISTILLED EMBEDDINGS ARE EASIER TO GENERATE

To verify that the distilled student is genuinely more diffusible, we compare the embeddings generated by ELF-B trained on the teacher and student embeddings in Fig. 6. By inspecting the generated samples at NFE 32, we find that the teacher sample contains more invalid and uncertain embeddings than the student, along with more grammatical mistakes.

We also test the denoising behavior of the diffusion trained on T5Gemma-2 vs. the student embeddings via round-trip denoising. Given the same sentence and the same added noise, we see that the teacher has a much larger embedding error after denoising, changing the sentence, while the student returns valid embeddings closer to the original sentence, with meaningful fluctuations.

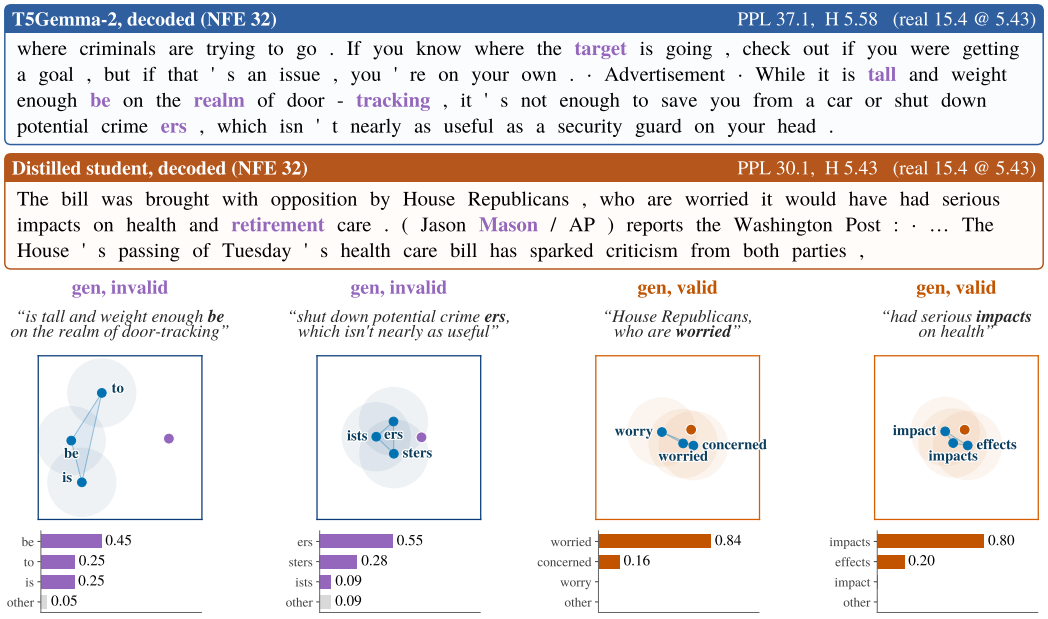


Figure 6: **Distilled embeddings are more diffusible with lower embedding error.** Same layout as Fig. 4. The teacher has more uncertain embeddings (in purple) which are mostly invalid, while the student's generated embeddings lie in the region formed by candidates that are close to each other. Over 64 sequences, the teacher has 142 uncertain positions per sequence, 109 of which are also invalid; the student 42 and 12; see App. A.1.

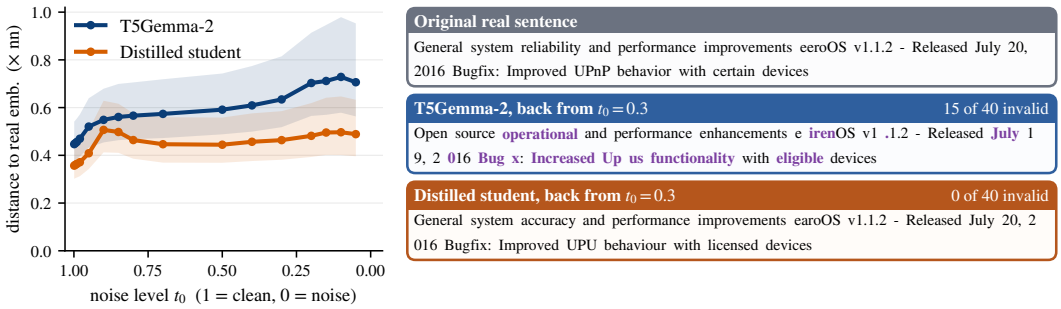


Figure 7: **Round-trip denoising from the same sentence.** *Left:* distance from each denoised embedding to the real embedding of the word it decodes to. *Right:* one sequence denoised from $t_0 = 0.3$; invalid tokens in purple.

5 Main Experiments

We evaluate the diffusibility of our embeddings by training the same ELF-B and ELF-M models on OWT-1024 (the full OWT partitioned into 1024-token sequences) and on LM1B (Chelba et al., 2013) (128-token sequences). For each embedding+diffusion, we sweep the self-conditioning scale and NFE, re-tokenize the samples with the GPT-2 tokenizer[3] and report the best Gen. PPL under GPT-2-Large at or above real-text entropy, and the mean and std are calculated over 3×1024 samples from 3 seeds.

[3]One problem is that, as different embeddings come with different tokenizers, the 1024-token sequences may not be directly comparable. So we re-tokenize everything with GPT-2 here. In App. B.3, we show that the tokenization effect is negligible and they are actually comparable.

Table 1: **Generation quality on OWT-1024.** Best Gen.PPL at or above real-text entropy over the sampling sweep; **bold** = best, underline = second best. †: cited. [a]: generation-time parameters, $\text{trunk}_{+\text{head}}$. [b]: MAUVE is protocol-sensitive and serves as a secondary metric. Gray rows fail to reach real-text entropy.

Method	Gen.#Params[a]	Gen.PPL(↓)	Entropy(↑)	MAUVE[b](↑)
Real Text	—	15.4±0.3	5.43±0.01	0.95±0.00
Base scale				
GPT-2-S (AR) (Radford et al., 2019)	85_{+39}M	34.1±0.8	5.45±0.02	0.88±0.02
Duo (Sahoo et al., 2025)	92_{+77}M	45.6±0.5	5.44±0.00	**0.95±0.02**
Aurora LM† (Liang et al., 2026)	130M	98.4±1.4	5.45±0.01	0.85±0.03
LDLM† (Meshchaninov et al., 2026b)	132_{+66}M	34.3±0.3	5.46±0.01	0.31±0.05
CoBit-S (Batzolis et al., 2026)	128_{+6}M	37.3±0.5	5.33±0.00	0.69±0.03
T5-small + ELF-B (Hu et al., 2026)	89_{+16}M	52.1±3.2	5.31±0.02	0.41±0.04
T5Gemma-2 + ELF-B	89_{+168}M	38.5±0.5	5.41±0.01	0.76±0.05
Student (ours) + ELF-B	89_{+168}M	**31.2±0.2**	5.44±0.00	0.86±0.01
Medium scale				
GPT-2-M (AR)	302_{+52}M	20.8±0.3	5.45±0.00	**0.91±0.01**
CoBit-M	453_{+9}M	31.3±1.1	5.43±0.01	0.49±0.03
T5-small + ELF-M	327_{+16}M	32.8±0.3	5.36±0.01	0.86±0.02
T5Gemma-2 + ELF-M	327_{+168}M	19.3±0.2	5.44±0.00	0.86±0.02
Student (ours) + ELF-M	327_{+168}M	**17.8±0.1**	5.45±0.01	0.89±0.01

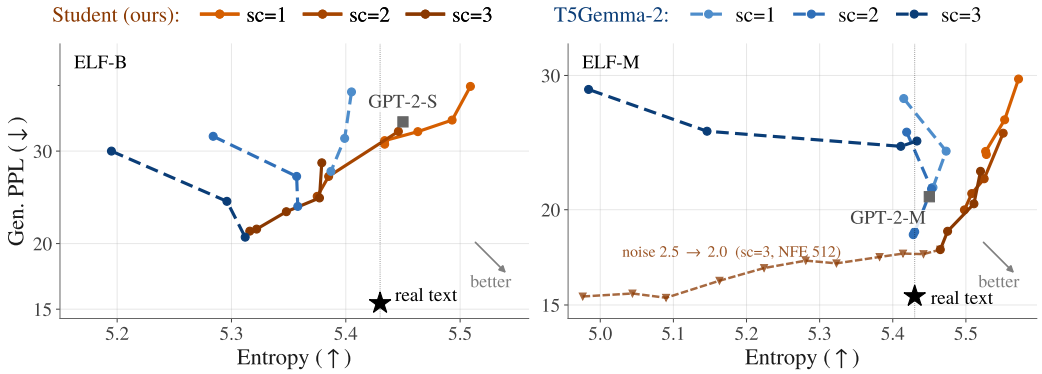


Figure 8: **Sampling ablation of ELF-B (left) and ELF-M (right) models.** Each curve stands for different NFEs (64–512) under the same self-conditioning. Our embedding consistently outperforms raw T5Gemma-2.

5.1 Generation Performance on OWT-1024

We compare our performance against previous state-of-the-art continuous DLMs. We include GPT-2-S/M as AR baselines, and also Duo (Sahoo et al., 2025) as a uniform-state discrete DLM representative. We find that ELF models trained on the T5Gemma-2 embeddings already reach the top tier, and our distilled embeddings further improve over that, outperforming previous methods on Gen. PPL at real-text entropy. Note that the pretrained encoders have seen far more text than the OWT-trained baselines and the original T5-small, so rows with different encoders are not fully controlled; the controlled comparison in Table 1 is T5Gemma-2 against the student.

Sampling Ablations We visualize the sampling sweeps in Fig. 8 that lead to Table 1. Across different sampling configurations, the student consistently reaches lower Gen. PPL at higher entropy than the T5Gemma-2 teacher. Notably, T5Gemma-2 collapses at extremely low NFEs (16–32) due to under-integration (see App. C for the collapsed samples), while our distilled embeddings are more stable.

5.2 Generation Performance on LM1B

We report the performance on LM1B (128-token sequences) to show that the improvement is not due to overfitting or familiarity with OWT data, but indeed a better geometry. The distilled student again performs well and surpasses the raw T5Gemma-2 embeddings.

Table 2: **Generation quality on LM1B.** Same protocol as Table 1. Duo has no public LM1B checkpoint. LM1B is a problematic corpus (Pynadath et al., 2026) and serves as a check rather than a comparison.

Method	Gen.PPL(↓)	Entropy(↑)	MAUVE(↑)
Real Text (LM1B)	52.7±0.6	4.29±0.01	0.95±0.01
AR-B (re-trained)	**37.7±0.5**	4.29±0.00	0.09±0.01
CoBit-S	59.4±0.6	4.31±0.01	—
LDLM†	63.0±0.5	4.37±0.00	0.91±0.02
T5Gemma-2 + ELF-B	60.6±0.3	4.32±0.01	0.36±0.00
Student (ours) + ELF-B	<u>50.5±0.5</u>	4.31±0.00	0.36±0.04

5.3 Efficiency

Scaling the embeddings is cheap for both training and sampling, as decomposed in Table 3. For training, the embeddings are frozen and take up a small part of the FLOPs and latency, and can also be cached; our distilled student halves the encoding cost of T5Gemma-2. For sampling, they affect only the final decoding step by the LM head, which is fast and performed once per sample. This is why LM head parameters are less important and are put in a subscript in Table 1. End-to-end, ELF-M produces a 1024-token sequence in 1.5 s at NFE 64 and 11.8 s at NFE 512; GPT-2-M takes 10.7 s with 1024 serial steps.

Table 3: **Cost of scaling the embeddings** per 1024-token sequence (one H100, bf16, batch size 1).

		Params	FLOPs	Latency
Training	Encoding, T5-small	19_{+16}M	0.04 T	1.5 ms
	Encoding, T5Gemma-2	100_{+168}M	0.21 T	3.4 ms
	Encoding, Student	50_{+168}M	0.10 T	2.1 ms
	ELF-B denoiser step (fwd+bwd)	89M	0.55 T	11.2 ms
	ELF-M denoiser step (fwd+bwd)	327M	2.0 T	30.6 ms
Sampling	ELF-B denoiser step	89M	0.18 T	11.8 ms
	ELF-M denoiser step	327M	0.67 T	23.0 ms
	Decoding, the denoiser pass	one denoiser step as above		
	Decoding, the 262k LM head	168M	0.34 T	6.3 ms

6 Conclusion

In this paper, we explore scaling the text embeddings that continuous DLMs diffuse on, and find that scaling genuinely improves diffusibility. We also understand diffusibility from the view of generating valid and decodable embeddings, and propose a distillation that further enhances the scaled embeddings. This gives a standard and reliable way to scale continuous DLMs from the embedding side, and shows their potential to co-evolve with the advancing foundation models.

Our results serve as a starting point for what this paper leaves open: few/one-step generation (given raw T5Gemma-2 under-integrates at low NFE), practical-scale models for real-world tasks such as QA, and adapting existing AR models directly into embeddings for DLMs (T5Gemma-2 is adapted from Gemma-3 (Team et al., 2025), and one would want to skip that step), which would make continuous DLMs a fast-sampling counterpart of AR models.

ETHICS STATEMENT

We use fully public corpora (OpenWebText, LM1B, WikiText-103) and models (T5Gemma-2, T5-small, ModernBERT). No human subjects or private data are involved. Our models generate unfiltered web-style text that can be false or biased; we use the samples only for evaluation.

REPRODUCIBILITY STATEMENT

Our continuous DLM code is adapted from the public ELF-pytorch implementation. We change only the embedding model and keep the training and sampling recipe (App. B.1). Distillation details are in App. B.2; App. B.3 gives the evaluation protocol, the sampling grid, and how each table row is chosen. Baselines are re-run with their authors' code and checkpoints, or cited where no checkpoint is public. Generated samples are in App. C.

DISCLOSURE OF AI USE

We use generative AI tools to implement methods (training, sampling, and evaluation codes), to help design and give feedback on experiments, to clean and reformat datasets, to search for additional literature, and to support qualitative analysis of generated samples. We have not used them to generate synthetic datasets, to formulate mathematical claims, to write proofs, or to assist with translation; the other required disclosures are not applicable. Additionally, we use generative AI tools to create and edit figures, to draft parts of the text (as suggestions; the final text is written by the authors). The research ideas, including scaling the embeddings and distilling them, originate from the human authors; AI tools are given direct, clear instructions and refined over multiple turns. All AI-assisted code is run and checked by the authors, and the results are re-checked against the external baselines. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in neural information processing systems*, 37:134614–134644, 2024.

Muhammad H Ashiq, Samanyu Arora, Abhinav N Harish, Ishaan Kharbanda, Hung Yun Tseng, and Grigorios G Chrysos. Why ddim hallucinates more than ddpm: A theoretical analysis of reverse dynamics. *arXiv preprint arXiv:2605.06831*, 2026.

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.

Georgios Batzolis, Mark Girolami, and Luca Ambrogioni. Cobit: Language modeling with bitstream diffusion. *arXiv preprint arXiv:2605.07013*, 2026.

Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025. URL `https://bfl.ai/research/representation-comparison`.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

Zhengdao Chen. On the interpolation effect of score smoothing in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 69869–69902, 2026.

Colin B Clement, Matthew Bierbaum, Kevin P O'Keeffe, and Alexander A Alemi. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*, 2019.

Jan De Leeuw and Patrick Mair. Multidimensional scaling using majorization: Smacof in r. *Journal of statistical software*, 31:1–30, 2009.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Empowering diffusion models on the embedding space for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4664–4683, 2024.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

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. In *International Conference on Learning Representations*, volume 2025, pp. 5046–5073, 2025.

Hongcan Guo, Qinyu Zhao, Yian Zhao, Shen Nie, Rui Zhu, Qiushan Guo, Feng Wang, Tao Yang, Hengshuang Zhao, Guoqiang Wei, et al. Continuous latent diffusion language model. *arXiv preprint arXiv:2605.06548*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

Jiaxiu Jiang, Jingjing Ren, Wenbo Li, Bo Wang, Haoze Sun, Yijun Yang, Jianhui Liu, Yanbing Zhang, Shenghe Zheng, Yuan Zhang, et al. Textldm: Language modeling with continuous latent diffusion. *arXiv preprint arXiv:2605.07748*, 2026.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Jean-Marie Lemercier, Tomas Geffner, Karsten Kreis, Morteza Mardani, Arash Vahdat, and Ante Jukić. Diladiff: Distilled latent-augmented diffusion for language modeling. *arXiv preprint arXiv:2605.23605*, 2026.

Na Li, Yuchen Jiao, Changxiao Cai, and Gen Li. Convergeflow: Language flow with provable convergence to token embeddings. *arXiv preprint arXiv:2608.23551*, 2026.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Jiajun Liang, Yucheng Liao, Yukang Cao, Jiazhe Wei, Ken Li, Wende Tan, Jiankun Zhang, ZY Cui, Jingkang Yang, Liucheng Guo, et al. Aurora-lm: Autoencoding unified representation for continuous-latent diffusion language modeling. *arXiv preprint arXiv:2608.02602*, 2026.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36:56998–57025, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Viacheslav Meshchaninov, Egor Chimbulatov, Alexander Shabalin, Aleksandr Abramov, and Dmitry Vetrov. Cosmos: Compressed and smooth latent space for text diffusion modeling. *Advances in Neural Information Processing Systems*, 38:14271–14299, 2026a.

Viacheslav Meshchaninov, Alexander Shabalin, Egor Chimbulatov, Nikita Gushchin, Ilya Koziev, Alexander Korotin, and Dmitry Vetrov. How to train your latent diffusion language model jointly with the latent space. *arXiv preprint arXiv:2605.07933*, 2026b.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

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. *arXiv preprint arXiv:2502.09992*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative frontiers: Why evaluation matters for diffusion language models. *arXiv preprint arXiv:2604.02718*, 2026.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *Proceedings of machine learning research*, 267: 52584, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

ByteDance Seed. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Alexander Shabalin, Simon Elistratov, Viacheslav Meshchaninov, Ildus Sadrtdinov, and Dmitry Vetrov. Why gaussian diffusion models fail on discrete data and how to prevent it? *arXiv preprint arXiv:2604.02028*, 2026.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, volume 2026, pp. 154506–154537, 2026.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

DiffusionGemma Team, Adrien Ali Taïga, James Assiene, Daniele Calandriello, Rahma Chaabouni, João Gante, Tamara von Glehn, Nate Keating, Chris Knutsen, Martin Kukla, et al. Diffusiongemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

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. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. *Advances in Neural Information Processing Systems*, 37:137432–137473, 2024.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *arXiv preprint arXiv:2603.05630*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in neural information processing systems*, 33:9422–9434, 2020.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *arXiv preprint arXiv:2504.06225*, 2025a.

Biao Zhang, Paul Suganthan, GaÃĢl Liu, Ilya Philippov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, et al. T5gemma 2: Seeing, reading, and understanding longer. *arXiv preprint arXiv:2512.14856*, 2025b.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. Planner: Generating diversified paragraph via latent language diffusion model. *Advances in Neural Information Processing Systems*, 36:80178–80190, 2023.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Cai Zhou, Chenxiao Yang, Yi Hu, Chenyu Wang, Chubin Zhang, Muhan Zhang, Lester Mackey, Tommi Jaakkola, Stephen Bates, and Dinghuai Zhang. Coevolutionary continuous discrete diffusion: Make your diffusion language model a latent reasoner. *arXiv preprint arXiv:2510.03206*, 2025.

A Additional Experiments

A.1 Statistical results on the generated embeddings

We generate 64 sequences of 1024 tokens per model with ELF-B at NFE 32 and keep every token position where the decoded top-1 is below 0.6: 9064 tokens for the teacher and 2664 for the student. For each, we measure the distance from the generated embedding to the nearest candidate word. The teacher's embeddings land far (mean 1.39 nn, median 1.38), the student's close (mean 0.82, median 0.81). Per sequence, the teacher has 142 uncertain positions, of which 109 are invalid; the student 42 and 12. Since we only inspect uncertain positions, these invalid counts are lower bounds.

Fig. 9 shows that a large embedding error genuinely means uncertain decoding: for the teacher, the median top-1 falls from 0.48 within the 0.91 nn range to 0.23 beyond it; for the student, it stays near 0.48 and 0.45. And the embedding error genuinely hurts the text quality: the 41 (out of 64) teacher sequences whose median distance is within 0.91 nn have Gen. PPL 18–106, and the two beyond 1.8 nn read above 2000; every student sequence has a median distance below 0.95 nn.

A.2 Candidates are robust to different decoders

We read the candidates of a token from the deployed decoding head, as it is the map from embeddings to words that the DLM uses, and the top-k are the words it would decode to. To check that the decoder choice does not drive the analysis, we run the pretrained T5Gemma-2 decoder on the

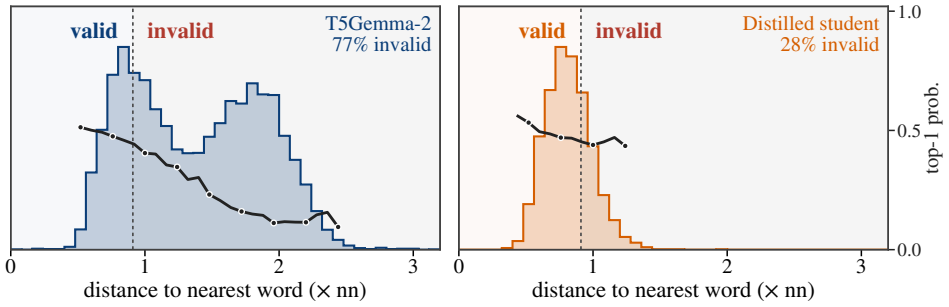


Figure 9: **Student's and teacher's generated embeddings.** Positions with top-1 below 0.6 from 64 sequences per model (NFE 32). Curve: the decoder's median top-1 per distance bin.

same sequences with teacher forcing. On real text, its top-3 shares at least one word with the head's at 99.3% of 764 positions (1.5 of 3 on average). On generated text, the two decoders are unsure at the same positions: the entropy of their predicted word distributions correlates at 0.72 over 3702 positions. The pretrained decoder is less certain overall (median entropy 6.3 nats), but it flags the same positions.

A.3 Ablation on distillation

Distillation objective We compare three training objectives (KL with soft labels, MSE with teacher embeddings, CE with input labels) for distillation and train the same ELF-B while keeping any other thing fixed. As shown in Fig. 10, the MSE student reaches an entropy of only 5.26, while the CE-student produces samples with even lower entropy (below 4.18). The raw teacher also stays below real-text entropy. In contrast, the student distilled from the teacher's soft decoder probabilities surpasses the teacher and achieves a Gen. PPL of 31.2 at a real-text entropy of 5.44. So the distillation targets (soft labels), rather than backpropagating gradients from the decoder, make the embeddings diffusible.

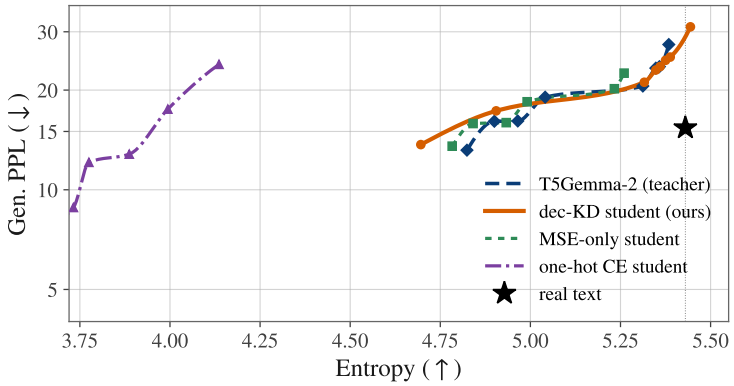


Figure 10: **Ablation on the distillation objective.** Gen. PPL–entropy curves on OWT-1024 with ELF-B. All models are evaluated over the same sampling grid with 256 samples. Among the compared objectives, only distillation from the teacher's soft decoder probabilities reaches the real-text entropy range.

Student depth effect We vary the student's depth from 3 to 18 layers and keep everything else fixed (Fig. 11), and report the GenPPL-entropy curves. Depth matters most near real-text entropy: at 5.05, students with 12–18 layers read 31–34 Gen. PPL, the 9-layer student 49, and the teacher 46; 3 and 6 layers fall behind. At lower entropy, the curves are within about 2 PPL. The 18-layer student is as deep as the teacher and still beats it, so the gain does not come from fewer layers. We use 9 layers to halve the encoding cost. A randomly initialized 9-layer student does not converge.

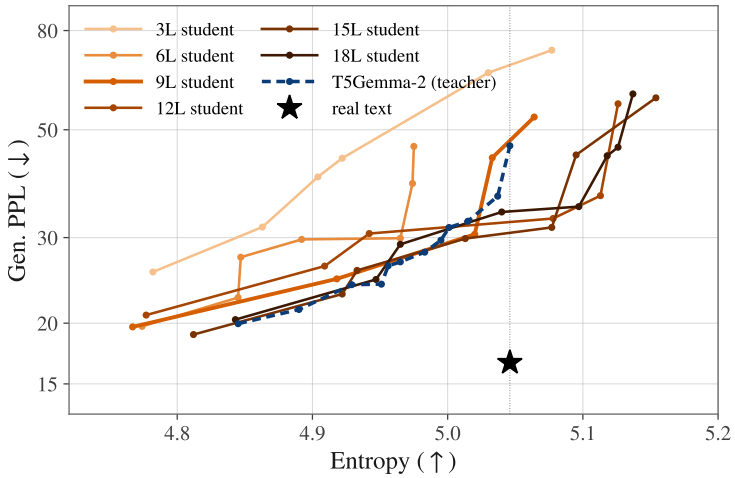


Figure 11: **Effect of student depth.** Gen. PPL–entropy curves for students with 3 to 18 layers and the T5Gemma-2 teacher on the 512-token OWT setting. All students use the same distillation and evaluation protocol. The vertical dotted line marks the entropy of real text, and the black star shows the real-text reference. Deeper students generally provide a better quality–diversity trade-off, although the improvement is not monotonic beyond 12 layers.

A.4 How distillation changes representation quality

Although the student is distilled only on OWT, it still encodes and decodes other text: top-1 token accuracy is 99.3 for both teacher and student on WikiText-103 (Merity et al., 2016), and 99.3 vs. 99.4 on held-out arXiv papers (Clement et al., 2019) (Table 4). Discriminative power drops a little. With the encoder frozen, a linear probe on the mean-pooled tokens scores lower on SST-2, STS-B and MRPC (ridge regression for STS-B). Therefore, distillation keeps token-level reconstruction and changes how the semantics are arranged.

Table 4: **Distillation trades discriminative power for diffusibility.** Reconstruction and probing results are reported on a 0–100 scale. Gray values indicate generation results below the real-text entropy.

Encoder (+ELF-M)	Generative Gen. PPL@Entropy (↓)	OOD reconstruction WikiText-103	OOD reconstruction arXiv	Discriminative SST-2	Discriminative STS-B	Discriminative MRPC
T5Gemma-2	19.3@5.44	99.3	99.3	89.4	71.7	71.6
Student (ours)	**17.8**@5.45	99.3	99.4	78.1	65.3	70.8
T5-small	32.8@5.36	100.0	100.0	84.4	74.3	74.8

B Experimental details

We build on a public PyTorch implementation of ELF.[4] Unless otherwise specified, we follow the original training and sampling configuration as in ELF.

B.1 Training Details

Main experiments. We pack OWT (Gokaslan et al., 2019) into 1024-token sequences with EOS between documents, and obtain 9.0M sequences (9.2B tokens) from the 7.8M documents. Following exactly ELF's design, we train ELF-B and ELF-M with x-prediction flow matching, logit-normal time sampling, and learned prefix tokens that carry the self-conditioning scale, for 5 epochs at batch size 512, using Muon optimizer with learning rate 2×10^{-3}, 2k warm-up steps, and EMA 0.9999;

[4] https://github.com/Ugness/ELF-pytorch

we evaluate the EMA weights. On LM1B, ELF-B trains on 7.9M 128-token sequences for 5 epochs with the same recipe.

Encoder comparison. To compare different pretrained embedding models at lower computational cost, we train ELF-B on a 25% subset of OWT using 512-token sequences. This setting is used for the encoder comparison in Fig. 2, including T5-small and T5-base (Raffel et al., 2020), T5Gemma-1 (Zhang et al., 2025a), T5Gemma-2 (Zhang et al., 2025b), and ModernBERT (Warner et al., 2025). We train on 2M OWT documents, truncated to 512 tokens (shorter ones padded), for 15 epochs (58.6k steps) with the same recipe. The real-text reference at this length is 16.6@5.05 (15.4@5.43 at 1024 tokens), so we compare encoders only within this setting.

B.2 Distillation details

We initialize the 9-layer student encoder using the even-indexed layers of the 18-layer T5Gemma-2 (Zhang et al., 2025b) teacher. Only the student encoder is trained, while the others (teacher encoder and decoder, vocabulary embedding table) are frozen. For each training sequence, the teacher decoder is teacher-forced on the teacher encoder outputs to produce a token distribution at each position. We retain the teacher's top-256 logits and renormalize them to form the soft target distribution. The student representations are passed through the same frozen decoder, and the student encoder is trained to minimize the token-wise KL divergence to the teacher distribution, implemented as cross-entropy over the retained indices, and we mask the padding positions.

We distill the student for 50k steps on the same packed 1024-token OWT sequences used to train the diffusion models. We use AdamW with a learning rate of 3×10^{-4}, 1k warm-up steps, and a batch size of 32 per GPU across 8 GPUs. We select the checkpoint with the lowest held-out distillation loss. Training takes 38.6 hours on 8 H100 GPUs. After distillation, we train a separate ELF model on the student embedding space using the identical architecture and training configuration as the one trained on the original T5Gemma-2 embeddings.

B.3 Evaluation Details

Metrics. We re-tokenize every generated and real-text sample using the GPT-2-Large tokenizer (Radford et al., 2019). We compute Gen. PPL with GPT-2-Large and unigram entropy under the same tokenization. The held-out OpenWebText (OWT) reference (Gokaslan et al., 2019) has a Gen. PPL of 15.4 and an entropy of 5.43. For MAUVE (Pillutla et al., 2021), we use the GPT-2-Large featurizer with 256-token features, 1024 generated samples, and a fixed held-out reference set of the same size. The real-text MAUVE value is computed between disjoint held-out splits and is below 1 because the metric is estimated from finite samples.

Sampling and operating-point selection. Unless otherwise specified, we use the SDE sampler adopted by ELF (Hu et al., 2026), with churn $\gamma = 1.5$ and noise scale 2.0, where NFE denotes the number of denoiser evaluations. For each combination of embedding space (T5Gemma-2 (Zhang et al., 2025b) or our distilled student) and diffusion model (ELF-B or ELF-M), we evaluate a shared grid of self-conditioning scales (sc) $\{1, 2, 3\}$ and NFEs $\{64, 128, 256, 512\}$ using 256 samples with seed 0, as shown in Fig. 8. Among the operating points that reach the real-text entropy of 5.43, we select the one with the lowest Gen. PPL. If no operating point reaches this value, we report the highest-entropy point and mark the result in gray. We then evaluate each selected operating point using three seeds with 1024 samples per seed and report the mean and standard deviation. The operating points reported in Table 1 are sc= 1/NFE= 256 for Student+ELF-B, sc= 1/NFE= 64 for T5Gemma-2+ELF-B, sc= 3/NFE= 512 for Student+ELF-M, and sc= 2/NFE= 256 for T5Gemma-2+ELF-M.

Baseline evaluation protocols. In Table 1 and Table 2, rows marked with † are taken from the corresponding papers under their original evaluation protocols and are therefore not directly comparable to results obtained under our protocol. Rows without † correspond either to our models or to public checkpoints that we resample using the authors' code and evaluate with 3×1024 samples under our protocol. We use $T = 512$ with nucleus sampling at $p = 0.94$ for Duo (Sahoo et al., 2025) and the released sampler for CoBit (Batzolis et al., 2026). Generation-time parameters are reported

as trunk$_{+\mathrm{head}}$, where the head is the token embedding table and output projection (for CoBit, its bit projections). Aurora LM's count is the authors' denoiser alone, since its autoencoder is not public.

GPT-2-S and GPT-2-M are sampled at temperatures 0.925 and 0.885, the temperatures that reach real-text entropy. On LM1B, every point of the OWT grid stays above real-text entropy (4.29), so both ELF-B models use sc$=$ 6/NFE$=$ 256. AR-B is a 12-layer GPT-2-style model trained from scratch on the same data (T5Gemma-2 tokens, 5 epochs), sampled at temperature 0.8. CoBit-S never reaches real-text entropy on OWT under its released sampler, so Table 1 shows its best Gen. PPL within 0.1 nats of it (37.3 ± 0.5 at 5.33). On LM1B we score CoBit-S under its authors' protocol (uncased WordPiece text, entropy over BERT ids, GPT-2-Large judge) and reproduce their number (59.4 ± 0.6 at 4.31; reported 59.8 ± 0.6); its MAUVE is left blank, since uncased text is not comparable to our cased references. LDLM rows are the authors' numbers (OWT-1024 at 256 NFE, LM1B at 512 NFE).

Tokenization One problem with the evaluation is that the encoders use different tokenizers (e.g., the T5Gemma family uses the 262k multilingual Gemma vocabulary, GPT-2 uses a 50k BPE), so a 1024-token sequence covers a different amount of text in each. Here, we verify that the improvements are not hacked by using different tokenizers or generating texts of longer effective lengths.

To make the methods comparable, we re-tokenize both generated and real text with the GPT-2-Large tokenizer before computing Gen. PPL and unigram entropy. At the sampling configurations used in the main evaluation (Table 1 and Fig. 8), the teacher and student samples contain median lengths of 1021 and 1027 GPT-2-Large tokens per window, respectively, compared with 1008 for real text. Their character counts are also similar (4.7k and 4.8k versus real 4.5k), and non-ASCII characters account for 0.3% in all three. We also find that truncating real windows by 5% changes unigram entropy by only 0.031, suggesting that the observed 1–2$\%$ difference in token length contributes < 0.01 to the reported entropy.

Table 5: **Sensitivity of MAUVE to feature length.** We vary only the maximum feature length while keeping the generated samples, reference samples, and GPT-2-Large featurizer unchanged. Increasing the feature length substantially lowers the scores of all models. We report three decimal places to distinguish the small scores and standard deviations in this analysis.

	MAUVE (↑)	
	256-token features	1024-token features
Real vs. real	.944	.924
GPT-2-S	$.878 \pm .020$	$.183 \pm .026$
T5-small + ELF-B	$.415 \pm .037$	$.012 \pm .004$
T5-small + ELF-M	$.858 \pm .017$	$.016 \pm .001$
T5Gemma-2 + ELF-B	$.758 \pm .052$	$.103 \pm .006$
Student + ELF-B	$.856 \pm .014$	$.111 \pm .009$
GPT-2-M	$.911 \pm .010$	$.169 \pm .015$
T5Gemma-2 + ELF-M	$.861 \pm .023$	$.098 \pm .006$
Student + ELF-M	$.889 \pm .013$	$.109 \pm .008$

B.4 Embedding-space Visualization Details

Each embedding panel in Figures 1, 4, and 6 visualizes a single token position. We show the generated embedding produced by the diffusion model together with the candidate embeddings associated with the decoder's top-3 predicted tokens at that position. All distances are first computed in the full 640-dimensional embedding space and then normalized by the median nearest-neighbor distance between real embeddings in that space, denoted by nn. This reference distance is 13.5 for the teacher and 12.4 for the student. A distance of $1\,\mathrm{nn}$ therefore represents the typical spacing between neighboring real embeddings, allowing teacher and student panels to be compared relative to the local scale of each space. The bars show the decoder's probabilities. Decoding is nonlinear, so the most probable word is not always the geometrically closest candidate.

We obtain the plotted coordinates using weighted metric multidimensional scaling with SMACOF (De Leeuw & Mair, 2009). In the dimension reduction, the distances between the generated em-

bedding and each candidate get weight 2.5, since these are the distances we want preserved. Across the displayed panels, the mapped generated-to-candidate distances differ from their full-dimensional values by at most 13%. We use a distance-preserving map rather than a linear projection such as PCA because the embedding space is still highly nonlinear, and the generated embeddings can lie far outside the subspace spanned by the candidate embeddings; a projection is then not meaningful.

C GENERATED EXAMPLES

We provide qualitative examples from OWT-1024 and LM1B under different sampling configurations. They illustrate the generation behavior of ELF (Hu et al., 2026) models trained on the raw T5Gemma-2 (Zhang et al., 2025b) and our distilled student embeddings. Notably, the T5Gemma-2 examples collapse at low NFEs such as 64, producing a funny language mix since the Gemma vocabulary is multilingual, while the distilled student examples are stable and fluent.

C.1 OWT-1024

T5Gemma-2 + ELF-M (sc 2/NFE 64) **seq. PPL 19.4, H 5.13**

"また、最近最近報告が公開明していたがロシアの戦争の Украинеのデデグ切されたされた中に CIAの兵員の情報をス Feしせるのではないのではないのでしょうか? 今まだ、具体具体情報を分究することがかないできません,が確かにこの可能性は非常に信頼だと思うと思います."： CIVIL自由の分野に範囲では、スかるける私たちの persoonlijke情報情報を機かりを採させるようなことで、国民の邦の政府に横略られてられてしまう悪悪になるになってになっててせなければことになるかも、固固擁られるcivil libertrechte、自由の表現と出版典権利の損レろられてられるようなような状況気がているというものの、今回当の日に始まりりにSupreme Court結論に決明にられて上がようになりました. Today, May 22, 1993, after years of legal battles, the Supreme Court finally reached decision on the legality of a bill that grants the president the power to interfereude with the work of Congress …"

T5Gemma-2 + ELF-M (sc 2/NFE 64) **seq. PPL 24.2, H 5.48**

"Пре統領人 Donald Trumpがが航旅行アメリカをにあきてきて 오는 июняでをつ選た次の pres政ンス nomineeににになる、報告した日曜新聞雑誌きました. 講し講談を設置かれたを約束明られてのイベントのを観われする将、処処する初めて度のス脳ンスニアになどを former統領者がの歴史歴史をスク選されてきました. リ歴知の ΔＰの創立された者が、国民の人間の derechos保護、民主化、経済成長、アメリカと世界に協相維持の目的を掲指っている。 ＪＰの、 ΔDの、ＵДC'とＵＰの総委員ニアを連携な協会で次の presbiedンス nomineeにを選せされようにを行いくので、想定は、発表の開催続いては former統領者のの過去以前の航旅行はアメリカのの方を付行ってっていよう.BEIJ, China, October 13 (Reuters) - The world's most loved bird and most endangered native animal, the Yangtze Bird …"

T5Gemma-2 + ELF-M (sc 2/NFE 256) **seq. PPL 23.0, H 5.46**

"A series of advanced radio telescopes have been studying the cold atmospheric surface of a star called Cepulus, which lies 1.4 billion light years from our sun. A traditional astronomical theory is that the star's cold atmosphere is a result of gravitational activity and radiation. But two University of Michigan astronomers, both co-authors of a new paper published online this week in the Proceedings of the National Acad Society, suggest that the star is made up of two different types of clouds. One of these clouds contains tiny membranes similar to those that attach to our own ears, a portion of our vision system. The other cloud contains a small number of smaller splles smaller than a baseball ball at the end of a baseball field. The findings published in the June issue of Astronomomy suggest the clouds may have been formed in response to radiation, but the star is very close to its parent star, which makes it even further away from Earth…"

T5Gemma-2 + ELF-M (sc 2/NFE 256) **seq. PPL 18.2, H 5.46**

"Decades of research have shown that deep-fracking has led to the release of toxic methane and carbon dioxide into the water system during the drilling processes used to explore a well's depth. The administration's proposed measure would ban methane, carbon dioxide and other components of oil and gas from appearing in drinking water for at least five years. It would also require state regulators to regulate toxic materials in the use of drinking water.
The nation's oil and gas industry generates about $1 trillion in revenue and costs about $400 billion a year. Deep-fracking releases more than 11,000 cancer-threataging chemicals and about 10,000 tons of carbon dioxide. Toxicive chemicals are linked to health risks, including major health problems such as heart disease, immune immune disease, diabetes, diabetes, leukemia and cancer..."

Student + ELF-M (sc 3/NFE 64) **seq. PPL 14.2, H 5.51**

"The crash took place at an intersection near East 8th Street and Virginia Avenue. Police and law enforcement officers with the Chicago CHP Unit located the driver driving the vehicle. The suspect was identified as 27-year-old Michael Pattiani. He is charged with misdemeanor use of carrying a firearm, drug trafficking, illegal possession of controlled drugs, marijuana trafficking, and possession of one or more amounts of marijuana, manufacturing and distribution of commercial controlled drugs. The charges range from at least one month to life in prison and are subject to a maximum fine of $100,000..."

Student + ELF-M (sc 3/NFE 256) **seq. PPL 11.8, H 5.47**

"It's also interesting to note that Sanders is probably not the only guy around when it comes to fighting global warming. The Clinton campaign hasn't dropped a single word on the issue of climate change to date, and the popularity of the Sanders campaign's "skip the wall" rhetoric makes it clear how little the campaign has actually done on the issue. In the wake of this election, the results of recent exit polls have shown that both sides of the political spectrum have realized the need to call out each other with bogus statements before they can even get to the polls.
If you're ready to read more from the unbossed and unbought Politicus team, sign up for our newsletter here! Email address: Leave this field empty if you're human:At least four people were killed in an airstrike in eastern Syria on Sunday, coming just days after the U.N. Security Council passed a resolution in support of ending the three-year civil war..."

Student + ELF-M (sc 3/NFE 256) **seq. PPL 11.6, H 5.42**

"In June, Sunpay drew fire when it announced a new technology, ShadowPass, which offered the ability to encrypt phone calls and text messages without requiring a password or consent. The revelations of ShadowPass's encryption techniques put several ISPs, such as AT&T Broadband Co, Verizon Wireless Co, Time Warner Co, T-Mobile Co and Sunpay, under the microscope for alleged fraud, identity theft and data leaking.
In October, Sunpay reached a settlement with the U.S. Justice Department in New York for violating secret orders from the FBI. The company argued that even though it had agreed to comply with enforcement and information-sharing regulations, the Justice Department went too far and ordered the company to pay settlement costs and damages. The FBI later alleged that Sunpay used similar encryption techniques to ShadowPass in order to share customer data with foreign intelligence agencies."

C.2 LM1B

T5Gemma-2 + ELF-B (sc 6/NFE 256) **PPL 32.5, H 4.29**

"The well-known Illinois senator spent more than 12 years of working with the White House under President George W. Bush from 1991-98 , was deputy economic adviser for President Bush and had a key role in the beginning of President Bush 's administration .A US military official says the Taliban killed four British soldiers in a car attack at a military base in southern Afghanistan .Sales for the year rose by 24.4 % 2009 after a 5.8 % rise in 2008 , but August sales figures were far lower than many economists 's expectations. . . "

T5Gemma-2 + ELF-B (sc 6/NFE 256) **PPL 60.3, H 4.31**

"Other improvements at TG include the Interactive Television and Video Matching System (IMC) , the Interactive Abreel Recall Retrieval System (IAC) for file matching technology , and the Integrated Video in Data Modification (IMP) system that allows the television staff to watch television shows two to three hours .State police spokesman Ron Parner said the police were still investigating Wright 's case and that the agency had no evidence of any wrongdoing .In London , the British Bank of England (BCC) said on Tuesday that the economy could lose 80 per cent of the global economy in the second quarter and that unemployment would continue to rise ."

Student + ELF-B (sc 6/NFE 256) **PPL 30.1, H 4.32**

"The cost-reduction plan will see not only reduce operating costs ; it is focused on increasing costs and increasing staff loyalty to reduce the overall costs .In a statement released Wednesday , the company said it would make a net loss of more than $ 100 billion under its bailout plan , but could not immediately say how much income would be paid from the sale .A third man has died in the crash of a transport truck in Bournemouth , a spokeswoman said .Mr Wood said that the deal would not be made by the time of this year 's end , but that it was not clear whether he would sell the shares. . . "

Student + ELF-B (sc 6/NFE 256) **PPL 50.4, H 4.36**

"The cost of the digital screens rose more than half last year , rising from £8m to £16m in 2007 .Arini Szymeva won a world-record 62 at women 's Beijing Olympics last year , and in 2008 she was awarded $ 69m in wins at the China International Tennis Tournament .There are a variety of occasions in which you feel like you 're getting love .LONDON , England (AP) - A man accused of killing a British military officer in Afghanistan has been sentenced to five years and six months , British police said Thursday ."