Abstract

Despite a growing interest in diffusion-based language models, existing work has not shown that these models can attain nontrivial likelihoods on standard language modeling benchmarks. In this work, we take the first steps towards closing the likelihood gap between autoregressive and diffusion-based language models, with the goal of building and releasing a diffusion model which outperforms a small but widely-known autoregressive model. We pursue this goal through algorithmic improvements, scaling laws, and increased compute. On the algorithmic front, we introduce several methodological improvements for the maximum-likelihood training of diffusion language models. We then study scaling laws for our diffusion models and find compute-optimal training regimes which differ substantially from autoregressive models. Using our methods and scaling analysis, we train and release Plaid 1B, a large diffusion language model which outperforms GPT-2 124M in likelihood on benchmark datasets and generates fluent samples in unconditional and zero-shot control settings.¹

1 Introduction

Large language models lie at the center of recent advances in artificial intelligence. Shared across nearly all such language models is a common recipe: learn a model that maximizes data likelihoods using an autoregressive, left-to-right factorization. Maximum-likelihood pretraining has been a remarkably successful paradigm, leading to models that perform well on a range of downstream tasks and display complex behaviors like in-context learning [1, 28].

Thus far, autoregressive modeling has been a core part of this process due to its computational efficiency and empirical performance. However, this choice carries drawbacks. Autoregressive models generate tokens one at a time, making it difficult to perform long-range planning or controllable generation [17, 19, 21]. In addition, certain sequence distributions may be fundamentally more difficult to model autoregressively [22].

Given the importance of language modeling, these potential drawbacks motivate us to explore alternatives to the autoregressive approach. As a promising candidate, we turn to continuous diffusion models [32, 12], which have achieved state-of-the-art results in image modeling [5, 30, 31]. In language, prior works on diffusion models exist [e.g. 21, 10, 6], but these optimize non-likelihood-based objectives. Without the ability to use standard likelihood-based benchmarks [25, 14, 26], it is difficult to say precisely how these models compare to autoregressive models (see Section 7 for a discussion). Somewhat concerningly, there is no work showing that it is possible for diffusion language models to achieve any nontrivial likelihoods on standard benchmarks.

In this work, we explore the limits of likelihood-based diffusion language models. Our goal is to train and release a diffusion model which achieves better likelihoods than GPT-2 124M [29], which we consider the smallest widely-adopted autoregressive model today. To achieve this goal, we first

¹We release our code and pretrained models at https://github.com/igul222/plaid.
Figure 1: Plaid models scale predictably across five orders of magnitude. Our largest model, Plaid 1B, outperforms GPT-2 124M in zero-shot likelihood (see Table 2).

Our contributions are as follows:

1. We explore the design space of likelihood-based diffusion language models and propose an algorithmic framework called Plaid. We validate the design choices of Plaid through compute-matched ablations.

2. We study the scaling laws of Plaid training. Our analysis shows that the log-likelihood of Plaid models improves predictably with more compute. We derive a recipe for compute-optimal training which differs substantially from the usual autoregressive rule.

3. We train and release Plaid 1B, a large diffusion language model pretrained on OpenWebText2 [7]. Plaid 1B outperforms GPT-2 124M in zero-shot likelihood across six standard benchmarks. We demonstrate Plaid 1B’s ability to perform fluent and controllable text generation.

2 Variational Diffusion Models for language

In this background section, we formally define continuous diffusion models over text sequences, adopting the Variational Diffusion Models (VDM) framework [18] which is a natural fit for likelihood-based training (see Karras et al. [16] for a survey on other formalisms). For brevity, we simplify some details in our exposition and refer the reader to Kingma et al. [18] for details.

Consistent with prior work (e.g. Li et al. [21]), our basic approach will be to map discrete text sequences into a continuous space with a token-wise embedding function and then construct a diffusion model on the embedded data.

2.1 Forward diffusion process

Consider a sequence of tokens \( x = [x^{(1)}, \ldots, x^{(L)}] \) drawn from the data distribution \( q(x) \). We transform \( x \) into a sequence \( \tilde{x} \) of embedding vectors using an invertible token-wise embedding function \( \text{Embed}(\cdot) \), such that \( \tilde{x}^{(i)} := \text{Embed}(x^{(i)}) \).

The forward process is a Markov chain over latent variables \( z_t \) from \( t = 0 \) to \( t = 1 \) which progressively adds Gaussian noise to \( \tilde{x} \). Let \( \sigma^2(t) \) be some monotonic function that specifies the total noise.
We can therefore parameterize the forward process distribution $q$ with $T$ discrete timesteps as

$$q(x, z) := q(x)q(z_0|x) \prod_{i=1}^{T} q(z_{i/T}|z_{(i-1)/T}) \tag{1}$$

where $q(z_0|x) := \mathcal{N}(\tilde{x}, \sigma^2(0))$ and $q(z_t|z_s) := \mathcal{N}(z_s, \sigma^2(t) - \sigma^2(s))$. It follows from this that $q(z_T|z_t, x)$ is also Gaussian, which will be useful later.

2.2 Reverse generative process

We can approximate the forward process distribution $q$ by a learned Markov reverse process where time runs backward from $t = 1$ to $t = 0$. The reverse process with $T$ timesteps is defined via the decomposition

$$p_\theta(x, z) := p(z_1) \left( \prod_{i=1}^{T} p_\theta(z_{(i-1)/T}|z_{i/T}) \right) p(x|z_0). \tag{2}$$

Let $z_{i}^{(i)}$ denote the portion of $z_i$ at sequence position $i$. Then we choose $p(z_1) := \mathcal{N}(0, \sigma^2(1/I))$ and $p(x|z_0) := \prod_{i} p(x(i)|z_{0}^{(i)})$ where $p(x(i)|z_{0}^{(i)}) \propto q(z_{0}^{(i)}|x_i)$. Recalling that $q(z_{(i-1)/T}|z_{i/T}, \tilde{x})$ is Gaussian, for the remaining factors we choose $p_\theta(z_{(i-1)/T}|z_{i/T}) := q(z_{(i-1)/T}|z_{i/T}, \tilde{x} = \tilde{x}_\theta(z_{i/T}))$ where $\tilde{x}_\theta(z_i)$ is a denoiser neural network that approximates $\mathbb{E}_q[\tilde{x}|z_i]$. Finally, our generative model is given by the marginal distribution $p_\theta(x) = \int_z p_\theta(x, z)$. If $\tilde{x}_\theta$ is optimal, then the forward and reverse processes express the same joint distribution as $\sigma^2(0) \to 0$, $\sigma^2(1) \to \infty$, and $T \to \infty$.

2.3 Likelihood bound

To optimize and evaluate the likelihood, we can write a variational lower bound (VLB) for the log-likelihood as

$$-\log p_\theta(x) \leq -\text{VLB}(x) := D_{KL}(q(z_1|x)||p(z_1)) + \mathbb{E}_{q(z_0|x)}[-\log p(x|z_0)] + \mathcal{L}_T \tag{3}$$

where

$$\mathcal{L}_T := \sum_{i=1}^{T} \mathbb{E}_{q(z_{i/T}|x)}[D_{KL}(q(z_{(i-1)/T}|z_{i/T}, \tilde{x})||p_\theta(z_{(i-1)/T}|z_{i/T}))]. \tag{4}$$

In the $T \to \infty$ limit, $\mathcal{L}_T$ simplifies to

$$\mathcal{L}_\infty = -\frac{1}{2} \mathbb{E}_{t \sim U[0,1], z_t \sim q(z_t|x)}[\text{SNR}'(t) \| \tilde{x} - \tilde{x}_\theta(z_t) \|_2^2] \tag{5}$$

where $\text{SNR}'(t) := \frac{d}{dt} \frac{1}{\sigma^2(t)}$. We use Monte-Carlo estimates of the resulting continuous-time likelihood bound to train and evaluate our model.

2.4 Learned noise schedule

A crucial hyperparameter in diffusion models is the noise schedule $\sigma^2(t)$, which specifies how much noise to add at each time in the diffusion process. In our setting, the VLB is differentiable with respect to $\sigma^2(t)$ via the reparameterization trick. Moreover, the VLB is invariant to the value of $\sigma^2(t)$ except at $t = 0$ and $t = 1$ in the continuous-time limit.

We can therefore parameterize $\sigma^2(t)$ as a scalar-to-scalar neural network and learn it by gradient descent. We train the endpoints $\sigma^2(0)$ and $\sigma^2(1)$ to maximize the VLB, and the schedule in between the endpoints to minimize the variance of the Monte-Carlo estimate of the VLB. Minimizing the loss variance is a proxy for minimizing the gradient covariance trace, which generally speeds up learning. See Kingma et al. [13] for further implementation details about this training procedure.

3 The Plaid framework

In this section, we present a series of algorithmic improvements to the basic setup described in Section 2. The result is a framework for diffusion language models which we refer to as Plaid (Perplexity-based LAnguage Inverse Diffusion).
3.1 Learned embeddings

In an autoregressive language model, the embedding operation is simply the first layer of the neural network and thus can be treated as just another part of the network. This is not true of embeddings in diffusion language models, which play a more fundamental role: they determine the order in which different tokens get generated. Tokens whose embeddings are far apart become distinguishable early in the reverse process, whereas nearby embeddings are distinguishable only later, at low noise levels.

Despite the importance of embeddings in diffusion language models, the loss functions used in prior work [21, 6] lead to ill-posed problems when optimized over $W_{\text{Embed}}$: for example, if our objective is $L_2$ reconstruction, then collapsing the embeddings by setting $W_{\text{Embed}} = 0$ and $\hat{x}_\theta(z_t) = 0$ yields a degenerate solution with zero loss. Prior work addresses this with workarounds like choosing $W_{\text{Embed}}$ by hand [3, 53] or using heuristic regularizers [21] or constraints [6].

In contrast, the Plaid loss function is a bound on the log-likelihood of the discrete data, which is a meaningful objective over both the model weights and embeddings. We therefore optimize the embedding matrix $W_{\text{Embed}}$ jointly with the rest of the model without additional constraints.

3.2 Categorical reparameterization

When optimally trained, $\hat{x}_\theta(z_t)$ learns to approximate a conditional expectation $E[\hat{x}|z_t]$ over sequences of word embeddings $\hat{x}$. At low noise levels, some or all of the embeddings in $\hat{x}$ are deterministic given $z_t$, so an optimal $\hat{x}_\theta(z_t)$ should output these exactly. However, doing so requires memorizing embedding vectors to high precision somewhere inside the model parameters, which is a poor use of capacity.

Instead of forcing the model to memorize the embedding vectors, we reparameterize $\hat{x}_\theta(z_t)$ as an average of embeddings weighted by a softmax over tokens. More formally, let $f_\theta(z_t)$ be a neural network which outputs logits and define $\hat{x}_\theta(z_t) := W_{\text{Embed}} \text{softmax}(f_\theta(z_t))$. We can interpret $f$ as learning a posterior over each discrete token $x^{(i)}$ given $z_t$. This relates to methods proposed in prior work, but these either require proxy objectives [21, 6] or consider image models [3].

3.3 Output prior

When we interpret $f_\theta$ as a posterior over tokens, the optimal value of $f_\theta(z_t)$ is $\log q(x^{(i)}|z_t) + Z$, which decomposes as $\log q(z_{1:t}^{(i)}|x^{(i)}) + \log q(x^{(i)}|z_{t}^{(\neq i)}) + Z$ where $z_{t}^{(\neq i)} := \{z_{t}^{(j)} : j \neq i\}$. We view the first term as a prior constraining the model’s predictions to those which are plausible given $z_t$, while the second term models relationships between different tokens.

To make it easier to model $f_\theta$, we compute the first term in closed form as the log-density of a Gaussian $\mathcal{N}(\hat{x}^{(i)}, \sigma^2(t)I)$ and add it to the output. This leaves the neural network with only the easier task of estimating $\log p(x^{(i)}|z_{t}^{(\neq i)})$. Empirically, we found it helpful to linearly anneal in this prior over the first 5000 steps of training.

3.4 Learned conditional likelihood

Recall that our loss function [3] includes a conditional likelihood term $\log p(x|z_0)$. We are free to choose $p$ however we wish, and in Section 2 we chose a position-wise factorial model $p(x|z_0) := \prod_i p(x^{(i)}|z_0^{(i)})$, with a simple fixed distribution for each factor. This choice is optimal for sufficiently small $\sigma^2(0)$, but using a more powerful model allows $\sigma^2(0)$ to take a larger value, effectively truncating the reverse process and therefore making it simpler to learn.

Here we leverage the fact that, after applying the categorical reparameterization (Section 3.2), our neural network $f_\theta(z_t)$ can be interpreted as learning the logits for $q(x^{(i)}|z_t)$ at all positions $i$. We therefore choose to keep $p(x|z_0)$ as a factorial model, but define each factor $p(x_i|z_0^{(i)})$ using the more powerful learned model $\text{softmax}(f_\theta^{(i)}(z_t))$.

Implementing this change naively requires two evaluations of $f_\theta$ for each minibatch example during training, corresponding to the two terms of $\mathcal{L}_\infty$ and $\log p_0(x|z_n)$. We instead split each minibatch,
Table 1: Compute-matched ablations of algorithmic components on OpenWebText2.

<table>
<thead>
<tr>
<th></th>
<th>NLL bound (val.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our full method</td>
<td>3.89</td>
</tr>
<tr>
<td>Our full method (0.5× compute)</td>
<td>4.01</td>
</tr>
<tr>
<td>No learned noise schedule</td>
<td>4.17</td>
</tr>
<tr>
<td>No learned embeddings</td>
<td>4.54</td>
</tr>
<tr>
<td>No categorical reparameterization</td>
<td>4.25</td>
</tr>
<tr>
<td>No output prior</td>
<td>3.95</td>
</tr>
<tr>
<td>No learned conditional likelihood</td>
<td>4.03</td>
</tr>
<tr>
<td>No self-conditioning</td>
<td>3.98</td>
</tr>
<tr>
<td>CDCD [6] (our reimplementation)</td>
<td>4.23</td>
</tr>
</tbody>
</table>

using some examples to compute $L_\infty$ and the rest to compute $\log p_\theta(x | z_0)$. We allocate examples between the two terms according to the ratio $\sqrt{\text{Var}(L_\infty) / \text{Var}(\log p_\theta(x | z_0))}$, where we compute the variances using running estimates of each term’s first and second moments. This minimizes the variance of the full loss [3].

3.5 Self-conditioning

Self-conditioning [3] is a technique which improve the performance of diffusion language models. The core idea is to reparameterize the denoiser $\hat{x}_\theta(z_t)$ as the fixed point $y_\infty$ of a recurrence $y_0 := 0, y_{i+1} := \hat{x}_\theta(z_t, y_i)$ where $\hat{x}_\theta$ is a neural network which now takes two inputs instead of one. During training, we approximate the fixed point $y_\infty$ by randomly unrolling the recurrence to either $y_1$ (with probability 0.75) or $y_2$ (otherwise). When we unroll to $y_2$ during training, we zero the gradients with respect to $y_1$, the noise schedule, and the embeddings. During held-out likelihood evaluation, we always unroll to $y_2$. During sampling, instead of solving the recurrence from scratch at each step of the diffusion chain, we compute $\hat{x}_\theta(z_1, 0)$ for the first step and $\hat{x}_\theta(z_t, \hat{x}_\theta(z_{t+(1/T)}, \ldots))$ for subsequent steps.

3.6 Other details

We perform all forward and backward computations in double precision except for the Transformer layers themselves, which happen in bfloat16 mixed precision. This comes at a negligible extra cost since the Transformer layers dominate the overall cost.

Architecture choices We condition $\hat{x}_\theta(z_t)$ on the timestep $t$ by adding a sinusoidal encoding of $t$ to the Transformer’s residual stream before the first layer. Before feeding $z_t$ into the Transformer, we rescale it by a factor of $\sqrt{1 + \sigma^2(t)}$ which makes each input dimension approximately unit-variance. Whereas autoregressive Transformers are relatively insensitive to aspect ratio [15], we find that Plaid performance increases significantly with Transformer depth up to about 16 layers. We also find that performance is sensitive to the choice of embedding dimension, with small values performing best. In all experiments, we use embedding dimension 16.

Stochastic sequence length Unlike autoregressive models, diffusion language models can only operate on sequences of exactly the same length as those seen during training. To enable our model to generalize to shorter sequence lengths, we truncate a small random subset of examples seen during training to random lengths. We observe that truncating even 3% of examples allows the model to generalize well across lengths without impacting full-length performance. Short-sequence performance does not improve substantially as we increase the number of truncated examples.

4 Ablation experiments

In this section, we validate different aspects of the Plaid framework through compute-matched ablation experiments.
4.1 Validating likelihood-based training

We take a likelihood-based approach in this work for multiple reasons: it has a principled interpretation, it simplifies training and evaluation, and it has yielded strong results in autoregressive models. Here, we validate that the log-likelihood objective can attain competitive sample quality through a human evaluation.

In diffusion models, the log-likelihood bound is an expectation over noise levels of a reconstruction loss weighted by a specific function of the noise level. In contrast, most prior work on diffusion models for language [21, 33] as well as images [12, 5] use heuristic weight schedules. Motivated by the intuition that human perception is more sensitive to coarse structure than fine details, these typically assign more weight to higher noise levels than the likelihood weight schedule.

We train three Plaid models: one with the likelihood weight schedule (“VLB”) and two with heuristic weight schedules (“Schedule A” and “Schedule B”) which we plot in Appendix B. The models are trained on a large dataset of short children’s stories which we constructed by finetuning GPT-J [34] on ROCStories [27]. Because learning embeddings is only straightforward when training against the likelihood bound, all models use fixed embeddings obtained from a previously-trained known-good model.

We repeatedly asked crowdworkers to choose from a pair of model samples, where one sample came from the likelihood-trained model and the other came from a heuristically-trained model. On average, crowdworkers preferred the likelihood-trained model over both alternatives: Weighting A’s win rate was 0.449 ($p = 0.001$, 95% CI [0.417, 0.482]) and Weighting B’s win rate was 0.457 ($p = 0.005$, 95% CI [0.425, 0.490]). Because we only consider two alternative weight schedules, we do not claim that the likelihood objective yields optimal sample quality, but our results suggest that it performs at least comparably to other choices.

4.2 Validating algorithmic components

Having validated our likelihood-based approach, we show in this section that each of the algorithmic components described in Section 3 lead to improved likelihoods in a compute-matched ablation study.

We train Plaid models on OpenWebText2 [7] and report their log-likelihood bounds on held-out data in Table 1. Our reference model (“full method”) is a $16 \times 384$ Transformer with 28M non-embedding parameters, trained for 92K steps at batch size 256 and sequence length 256, corresponding to $1.12 \times 10^{18}$ non-embedding FLOPs. For each ablation model, we stay as close to this configuration as possible while preserving the number of non-embedding FLOPs (we exclude FLOPs from the embedding and output projections because these become negligible at large scale). We observe that ablating each of the components described in Section 3 results in a worse log-likelihood. As a comparison point, we also train a model at half the compute budget ($5.6 \times 10^{17}$ FLOPs) by halving the model size. See Appendix C for more training details.

Finally, as a comparison to prior work, we reimplement CDCD [6], train it following the same configuration, and report its log-likelihood. We follow the authors’ description as faithfully as possible except for the noise schedule endpoints, embedding dimension, and embedding weight initialization, which we tune to maximize log-likelihood. We observe in Table 1 that even the half-compute-budget version of Plaid surpasses our CDCD implementation in likelihood. Note that CDCD was not developed as a likelihood-based model, and the lack of a public implementation means that there are most likely differences between our implementation and the original.

5 Scaling laws for Plaid

Having developed an algorithmic framework for diffusion language models, we now study its scaling properties in order to guide large-scale training of Plaid models. In the case of autoregressive models, the work of Kaplan et al. [15] demonstrates that model log-likelihoods follow a log-linear scaling law: across many orders of magnitude, training with more compute predictably improves likelihood. Using these scaling laws, Kaplan et al. [15] and Hoffmann et al. [13] accurately predict the optimal model size as a function of the given compute budget across many orders of magnitude. Both results together enable effective large-scale training. In this section, we experimentally determine them for Plaid models.
Figure 2: Plaid models improve with compute at a similar rate to autoregressive models, but Plaid is less efficient by a constant factor of 64×.

Figure 3: Compute-optimal Plaid models should be 4× smaller (and trained for 4× longer) than compute-optimal autoregressive models.

5.1 Methodology

Our main experimental method will be an IsoFLOP analysis [13]. We first fix a set of FLOP budgets \( \{C_1, \ldots, C_K\} \). For each budget \( C_i \), we train models with different sizes \( \{N_{C,1}, \ldots, N_{C,M}\} \) and perform a quadratic fit of the loss \( L \) to \( \log N \). We plot all the data along with quadratic fits in Appendix D and find that the fits approximate the data well.

The minimum of the quadratic gives us the compute-optimal loss \( L_{C_i}^* \) and corresponding model size \( N_{C_i}^* \) for that budget.

Given the compute-optimal loss \( L_{C_i}^* \) for each FLOP budget \( C_i \), we fit the parameters of a loss scaling law

\[
\min_{\alpha, \beta} \sum_i (\log(L_{C_i}^*) - \beta \log(C_i) - \alpha)^2
\]

which can then be used to predict the compute-optimal loss as \( L^*(C) = \alpha C^\beta \). We also fit a parameter scaling law \( N^*(C) \) in the same fashion from the model sizes \( N_{C_i}^* \).

We perform IsoFLOP analyses for both Plaid and autoregressive models in order to compare the results. We choose compute budgets log-uniformly between \( 10^{16} \) and \( 10^{19} \) FLOPs and corresponding model sizes heuristically. We choose learning rates using \( \mu \text{Transfer} \) [15] and batch sizes, weight decays, and aspect ratios by well-tuned heuristics. When computing FLOPs, we exclude FLOPs from the embedding layers and output projections. This enables us to use much smaller compute budgets than Hoffmann et al. [13], but it causes our autoregressive scaling law to differ slightly from theirs. We consider this acceptable since we are mainly interested in the differences between our autoregressive and Plaid scaling laws.

5.2 Loss improves predictably with compute

We plot both of our scaling laws in Figure 2. Our first finding is that over many orders of magnitude, the compute-optimal log-likelihood of Plaid models closely matches a power law function of the compute. Surprisingly, we find that the slopes of both the autoregressive and diffusion scaling laws are almost exactly the same. These results validate Plaid’s scalability and suggest that we can obtain strong improvements by training at larger scale.

Regardless of scale, Plaid models require a constant factor of about 64× more compute to match their autoregressive equivalents. While this factor is large, our work represents the very first attempt at efficient diffusion model training and focused engineering effort on constant-factor improvements to diffusion models may enable them to perform similarly to autoregressive models in the future.
Table 2: Plaid 1B outperforms GPT-2 124M in zero-shot likelihood across six benchmark datasets from Radford et al. [29]. Our GPT-2 numbers differ from the originals due to striding and detokenization (see Section 6.1).

<table>
<thead>
<tr>
<th></th>
<th>PTB (PPL)</th>
<th>enwik8 (BPC)</th>
<th>text8 (BPC)</th>
<th>WikiText2 (PPL)</th>
<th>WikiText103 (PPL)</th>
<th>1BW (PPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plaid 1B (ours)</td>
<td>74.33</td>
<td>1.18</td>
<td>1.12</td>
<td>29.42</td>
<td>28.28</td>
<td>77.64</td>
</tr>
<tr>
<td>GPT-2 124M</td>
<td>87.97</td>
<td>1.24</td>
<td>1.22</td>
<td>35.01</td>
<td>35.92</td>
<td>87.85</td>
</tr>
<tr>
<td>GPT-2 345M</td>
<td>64.92</td>
<td>1.09</td>
<td>1.11</td>
<td>26.80</td>
<td>26.13</td>
<td>67.34</td>
</tr>
<tr>
<td>GPT-2 762M</td>
<td>53.42</td>
<td>1.04</td>
<td>1.06</td>
<td>23.30</td>
<td>22.24</td>
<td>59.48</td>
</tr>
<tr>
<td>GPT-2 1.5B</td>
<td>47.59</td>
<td>1.00</td>
<td>1.02</td>
<td>21.33</td>
<td>20.13</td>
<td>54.09</td>
</tr>
</tbody>
</table>

Table 3: Chosen unconditional samples from Plaid 1B demonstrate fluent syntax and long-range coherence. See Appendix E for un-picked random samples.

New research rolled out at an annual scientist meeting finds that the industry will need to recover between 4,000 and 10,000 tons every year of fracked and produced oil and gas fossil reserves in order to do that, according to an analysis done by lead author Dr. Ernesto Monteiro of the University of Alberta in Canada. The eye-watering figure represents the total amount of oil and gas – shale and natural gas produced, extracted and sold – will likely need to be recovered in coming years to meet the carbon mitigation goals.

The Barcelona Golf Course doesn’t look like a golf course, but it is an oasis for gardening in this busy city. It’s a massive course spread over 120 acres with fairways that are split right down the middle, unlike the designs on most golf courses. The course enjoys the stunning view of the skyline above it. A giant oak tree almost 40 meters in diameter serves as one main highlight to the golf course’s design.

The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

The team uprooted the previously existing Pérez Tree, to make room for the new trees to complement Gillet Park.

5.3 Compute-optimal training recipe

Our next goal is to understand how to optimally use a given compute budget $C$ to maximize the held-out likelihood of a model. Specifically, we must choose between training a large model for fewer iterations or training a small model for longer. For this, we leverage our parameter scaling law $N^*(C)$ which predicts the optimal model size given a compute budget.

We plot both of our parameter scaling laws in Figure 3 and again find that the trends have nearly the same slope but differ by a constant factor. Specifically, compute-optimal Plaid models should be about $4 \times$ smaller (and therefore trained for $4 \times$ longer) than compute-optimal autoregressive models. The large gap in compute-optimal settings suggests that selecting model sizes based on existing scaling laws [15, 13], which were developed for autoregressive models, could incur a substantial loss in the effective compute budget.

6 Plaid 1B

To demonstrate the scalability of Plaid models and achieve our goal of outperforming an autoregressive model in likelihoods, we train, evaluate, and release a large Plaid model called Plaid 1B. Plaid 1B is a Transformer-based Plaid model with 1.3B parameters, trained for 314B tokens on OpenWebText2 [7]. In total, Plaid 1B was trained for $2.5 \times 10^{21}$ FLOPs, which to our knowledge equals the largest purely diffusion-based language model trained in prior work [6]. We give further training details in Appendix C.

6.1 Likelihood evaluation

We evaluate Plaid 1B’s likelihood in a zero-shot setting on a suite of six benchmark datasets originally used in Radford et al. [29]: Penn Treebank [25], enwik8 and text8 [14], WikiText2 and WikiText103 [26], and the One Billion Word corpus [2].
Table 4: Chosen conditional samples from Plaid 1B in different zero-shot control settings. Highlighted spans are prompts. See Appendix E for un-picked random samples.

Prefix completion:

Generative models of text are very versatile: they can be used as a data classification model and also incorporated into multiple data processing engines. In this article, we present two new neural memory models capable of processing terabytes of data and the neural networks and computational techniques that are used in those models.

Infilling:

A year ago in Paris, prior to the tournament, I went to Elijah's to eat and get drunk. Everyone in the venue was seventeen. I was there for a few minutes and then I went back to the event. Wow, what a great day! So relaxed and too happy. I do not think I was always like that.

Token-level weights (5 \times weight on “law”):

Let’s talk about law and medicine.

In her dissent, Justice Ron Sen, a veteran administrative law judge, points out that the decision "ignores the fact that the original separation agreement was reached by binding arbitration" that responded to "the legitimate ethical concerns of the university administration," which is what lies "at the heart of law and medicine."

Token-level weights (5 \times weight on “medicine”):

Let’s talk about law and medicine.

In part because of advancements in technology, personal information about medical and drug use is spreading. Healthcare professionals across the nation rely on this personal data to make decisions about drug prescriptions and clinical trials and monitor people at immediate risk of serious or chronic diseases.

Lexical constraints (“Donald” anywhere):

Also facing legal challenges is Donald Trump’s executive order banning immigration from seven Muslim-majority countries that is facing a temporary halt, with nothing scheduled to go into effect. Two federal judges have ruled that such an order violates the establishment clause.

Composition and negation (“Donald” anywhere and “Trump” nowhere):

A month later, with little time to spare, the government hired Donald V. Davis, a former senior aide to Senator Tom Mondale of Minnesota and former Chief Security Operations Officer at the White House, to lead tactical centers.

Radford et al. [29] use sliding windows of size 32 in their likelihood computation. As a non-autoregressive model, Plaid doesn’t support sliding-window likelihood evaluations, so we use non-overlapping 1024-token sequences when computing likelihoods. Following Radford et al. [29], we use heuristic invertible detokenizers for PTB, 1BW, and WikiText to minimize the effect of tokenization artifacts on the perplexity results. For a fair comparison, we also recompute GPT-2 likelihoods using the same protocol, resulting in different numbers than Radford et al. [29].

In Table 2 we observe that Plaid 1B consistently outperforms the 124M parameter GPT-2 model, demonstrating that diffusion models are capable of scaling to perplexities on par with a small modern autoregressive model.

6.2 Unconditional samples

We generate from Plaid 1B by starting from \( z_1 \sim \mathcal{N}(0, \sigma^2/\tau) \), performing ancestral sampling of \( p(z_{t-1}/z_t) \) for \( T = 4096 \) steps, and finally \( \operatorname{argmax} p_\theta(x|z_0) \). Following Dieleman et al. [6], we sample using a score temperature of \( \tau = 0.9 \), which in our formulation corresponds to adding \( \frac{1-\tau}{\tau}(\hat{x}_\theta(z_t) - z_t) \) to \( \hat{x}_\theta(z_t) \) at each step.

We generate unconditional samples with sequence length 1024 and present chosen samples in Table 3.

We observe that the model is capable of generating fluent text and remaining on-topic over several hundred words. We provide random un-picked samples in Appendix E.

6.3 Zero-shot control

Although Plaid models are trained in a purely unconditional fashion, we present a zero-shot control technique called token guidance which allows us to implement a number of conditioning structures at generation time. We begin with classifier guidance, a technique which allows diffusion models to generate samples conditioned on an arbitrary attribute \( y \). Classifier guidance first trains a probabilistic classifier \( p(y|z_t) \) of \( y \) given noisy latents \( z_t \), and then biases the diffusion model’s sampling steps by

\[
\hat{x}_\theta(z_t) = \frac{1}{\tau}(\hat{x}_\theta(z_t) - z_t) + \frac{1-\tau}{\tau}y_t,
\]

where \( y_t \) is the target attribute value. We choose the splitting boundaries using the Plaid tokenizer, yielding sequences slightly shorter than 1024 tokens under the GPT-2 tokenizer.
a guidance term derived from the gradient of the classifier probability $\nabla_z \log p(y|z_t)$. Now, recall from Section 3.2 that our denoiser $\hat{x}_\theta(z_t)$ is parameterized in terms of a model $f_\theta(z_t)$ which learns the distribution over the token $x^{(i)}$ at each position $i$ given $z_t$. We can therefore implement many different conditioning structures via classifier guidance on probabilities derived from $f_\theta$ itself. We give a few examples:

**Conditioning on a span:** We perform guidance on the joint probability of the span under the factorial model $p(x^{(a:b)}|z_t) \propto \prod_{i=a}^b p(x^{(i)}|z_t)$, where $f_\theta$ estimates each factor in the product. This lets us implement prefix completion and infilling as special cases. **Lexical constraints:** In order to condition on the presence of a token without specifying its location, we perform guidance on the token’s probability under the unigram distribution $p(x^{(any)}|z_t) \propto \sum_i p(x^{(i)}|z_t)$, where $f_\theta$ estimates each term in the sum. **Token-level weights:** We can emphasize a specific conditioning token by multiplying the corresponding guidance term by a scalar weight. **Negation:** We condition on the negation of an attribute $y$ by performing guidance on the complement probability $1 - p(y|z_t)$.

Using Plaid 1B and token guidance, we generate samples under various zero-shot control settings. We present chosen samples in Table 4 and random samples in Appendix E. Despite being trained unconditionally, Plaid 1B is able to follow diverse conditioning structures.

7 Related work

We contribute to a growing body of work on diffusion-based language models [21, 3, 9, 33, 10, 6, 8, 24, 37, 23, 36, 11]. Our biggest departure from those works is that we aim for strong likelihood performance, which to our knowledge has not been attempted in any prior work except for an appendix result from Li et al. [21]. We therefore benchmark against well-known autoregressive models instead of prior diffusion language models.

The work most comparable to ours is CDCD [6], which is also a strong general-purpose diffusion language model. However, without the ability to use standard likelihood-based benchmarks [25, 14, 26], it is difficult to say precisely where CDCD stands in comparison to autoregressive models: in every result, either CDCD underperforms the autoregressive baseline, or the evaluation metric saturates and lacks the statistical power to distinguish the models. Many of the other works above share similar difficulties. In contrast, our likelihood-based approach enables unambiguous comparisons to widely-known models.

Other diffusion language model works consider more constrained settings like controllable generation [21] or sequence-to-sequence tasks [3, 9, 8, 37, 23, 36], or propose hybrid approaches involving pretrained autoregressive models [24, 11]. Particularly, in concurrent work, Han et al. [11] finetune an OPT 13B [38] model into a hybrid model which is autoregressive over 25-token blocks and uses diffusion within blocks. Compared to their work, we focus on the more general setting of training a fully diffusion-based language model from scratch.

Finally, our work builds on recent advances in diffusion models for images [32, 12, 30, 18, 5]. Most notably, we adopt the framework of Variational Diffusion Models [18] and extend it to language modeling.

8 Conclusion

In this work, we have taken the first steps toward a competitive likelihood-based diffusion language model. We built Plaid 1B, which matches GPT-2 124M in likelihood by combining several algorithmic improvements and a scaling law analysis. Our ablations show that maximizing likelihood does not substantially harm sample quality, and we show samples are fluent in both unconditional and zero-shot conditional settings. Despite this progress, substantial work remains: Plaid narrows the compute-efficiency gap between diffusion and autoregressive language models to $64 \times$, and we view this gap as a tractable and exciting open problem that may be addressed with further research.
Acknowledgements

We thank Sander Dieleman, Lisa Li, and John Thickstun for valuable conversations. This work was supported by an Open Philanthropy grant.

References


A Additional experiments

**Plaid 1B Sample Quality** We compare unconditional samples from Plaid 1B and GPT-2 through a human study. We generate samples of length 128 from both GPT-2 and Plaid 1B ($n = 1023$ samples each), and repeatedly ask Mechanical Turk crowdworkers to choose the most coherent sample from a pair (one GPT-2, one Plaid, blinded and randomly ordered). The crowdworkers found the models comparable, with the 95% CI for the win rate of Plaid ranging from 0.47 to 0.55, with a mean win rate of 0.51.

**Sampling Timesteps** When generating samples, we attempt to approximate the infinite-timestep use a naive sampler (vanilla ancestral sampling) paired with a very large number of timesteps than necessary (4096) to approximate the continuous-time limit. To validate that our choice of 4096 steps is sufficient to approximate the continuous-time limit (with respect to human judgement), we run a human binary choice preference study of unconditional Plaid 1B samples generated with 2048 and 4096 steps ($n = 1024$ samples each). We find that crowdworkers do not strongly prefer either choice, with the 4096-step samples having a win rate of 0.51 (CI: [0.47, 0.53]).

B VLB and heuristic weight schedules

![Figure 4: VLB weight schedule and heuristic weight schedules used in ablation experiments.](image)

C Experiment details

C.1 Dataset

Unless otherwise noted, all models in this work are trained on a subset of OpenWebText2 [7] which we filter to remove documents labeled as non-English. The data is tokenized using a 32K-token BPE tokenizer which we train on the OpenWebText2 training split.

C.2 Architecture

We use standard pre-activation Transformers models with RMSNorm normalization layers and GeLU nonlinearities throughout. Unless otherwise noted, all Plaid models have 16 Transformer layers, which we found to be approximately optimal for our scale. We scale autoregressive model depth approximately following [19]. For efficiency, our implementation uses FlashAttention [4] and other fused kernels wherever applicable. We train at sequence length 256 for all experiments except Plaid 1B.
C.3 Optimization

We optimize all models using AdamW with parameter-specific learning rates derived by $\mu$Transfer \cite{35} based on a learning rate of $1.4 \times 10^{-3}$ at width 256. Each parameter’s weight decay is set to $4 \times 10^{-5}$ where $\eta$ is that parameter’s learning rate. We use a linear warmup on the learning rate and weight decay over the first 2500 steps, followed by a linear decay to zero over training. We train at batch size 256 for algorithm ablations and 128 for scaling law experiments. All of our small runs take less than 24 hours on a single A100. We perform all forward and backward computations in double precision except for the Transformer layers themselves, which happen in bfloat16 mixed precision. This comes at a negligible extra cost since the Transformer layers dominate the overall cost. Our learning rate and precision choices are optional: when implemented carefully, our method trains stably and performs well when single-precision floats and a single learning rate for all parameters is used.

C.4 Plaid 1B training

We increase the base $\mu$Transfer learning rate to $2 \times 10^{-3}$ (at width 256). The denoiser network is a Transformer with 24 layers of width 2048 and a vocabulary size of 32K tokens, for a total of 1.3B parameters. We train for 1.2M steps at batch size 256 and sequence length 1024, for a total of 314B tokens. All other details are as written above. Training took 30 days on 8 A100s.

C.5 User studies

The overall experimental design follows a blinded randomized binary choice experiment. The details are as follows: We recruited crowd workers using Amazon Mechanical Turk (selection criteria: US location, 95% HIT approval rate, >1000 HITs approved). Workers were shown two random samples in random order and given the following prompt: “Given two short pieces of text, decide which is more coherent overall (i.e. has the fewest grammar mistakes and makes the most sense).” Workers were paid $0.15 per task, which we estimated to take less than 30 seconds on average.

D IsoFLOP profiles

![IsoFLOP profiles](image)

Figure 5: IsoFLOP profiles for autoregressive models (left) and diffusion models (right).
tends to create a ripple effect in the economy, and economist Kerin Milligan said it’s probably a testament to how low the economy and consumers are, particularly at the low end. “It’s probably due to things like investment slowing in the natural resource sector, weakness in stock buybacks in the corporate sector, or just other things that many economists are concerned about.”

Milligan said it’s particularly concerning how low mortgage rates, and even lower interest rates, continue below advertised rates. “It’s difficult to be selling at rates well below their 10-year average, so I think those effects do come back into the economy as well, and I’d wonder if it’s pulling down consumption and investment activity.”

Consumers and businesses are really looking for more and more type of stimulus, and they’re having a difficult time selling at rates well below their 10-year average, so I think those effects do come back into the economy as well, and I’d wonder if it’s pulling down consumption and investment activity. 

Responding to the incident, Ariana Grande tweeted:

‘I was at @Meteze show shock at the “serious” incident, Sky News reports.

JUST IN: ‘The Queen was targeted’: UK Police offer exact timings

The driver, arrested on suspicion of driving while working and driving

A man has been arrested for allegedly crashing into a group of three children walking

He said those predictions are becoming more consistent as well, with a reported saying from economist

He said: “It’s completely bonkers that

Scottish Labour Shadow Justice secretary Michael Russell said the fact officers

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that

The current pay structure was introduced in the late 1990s to compensate officers in frontline

He said: “It’s completely bonkers that

The Scottish Government said in 2017 the police service had more than 23,000 officers, a third of whom

He said: “It’s completely bonkers that
zone over the last four years.\n\n"Why have ministers allowed higher officer salaries in the police while they have delivered a decade of devastating budget cuts to the NHS?\n\nThe Scottish Government showed how finance and good colleagues will be putting a concerted effort into offsetting some of these increased costs, instead of deciding the way to save the workforce is a cutters field.\n\n"Nicola Sturgeon, the Scottish First Minister, recently said: "Front line firefighters and police officers are asked to endure some of the most distressing things on the basis that the average police officer pay rose to £43,085 in 2019, after a two-year increase, the AMPO said.\n\n"The rise in law enforcement is seen to have come about by an ageing population and the overseas population and the deepening of recession job cuts in the Scottish public sector.\n\nIt has been returned to the Pottona Police Department Monday afternoon. She was taken from a home in the 3400 block of Ponderosa Avenue near Fairfax Boulevard and Ligea Street.\n\nPolice say it is believed the child, who is staying with her family, was taken from her home between July 20 to July 22. Forensic testing will likely be performed later this week.\n\n"It was an odd fact that made many left-wing activists vehemently oppose the nomination of Brett Kavanaugh to the U.S. Supreme Court. On his yearbook page was a subtitle entitled "Don't Let Them F***** Your Body Forever," a reference to the oral contraceptive.\n\nof carbon emissions needed to meet the 2 degree Celsius target set forth in the last assessment of the Intergovernmental Panel on Climate Change (IPCC) in 2013. New research rolled out at an annual scientist meeting finds that the industry will need to recover between 4,000 and 10,000 tms every year of fracked and produced oil and gas reserves in order to do that, according to an analysis done by lead author Dr. Ernesto Monteiro of the University of Alberta in Canada.\n\nThe eye-watering figure represents the total amount of oil and gas - shale and natural gas produced, extracted and sold - will likely need to be recovered in coming years to meet the carbon mitigation goal.\n\n"The benefits of carbon capture increase with the amount being injected," Monteiro wrote in an academic journal prepared for the annual meeting of the National Academies of Sciences which 2 the week.\n\n"While the theoretical ability to limit 2 C warming with fossil fuel carbon budgets is obvious, such action is not taken by the oil and gas industry. Industry groups led by RBC openly seek to capture 38 million tons of carbon dioxide emissions and transfer a portion of the back to the oil and gas industry as revenues.\n\n"Comparing natural gas reserves under production with proven oil reserves remaining reveals one potential implication, which is the additional natural gas production that would be required to fully satisfy the 2100 oil supply balance," the paper explained.\n\nFrom where I impeach, McConnell announced that "we have the votes to convict the president."\n\nOn Wednesday, after the House Judicary Committee drew up articles of impeachment, McConnell announced that "we have the votes to convict the president."\n\n\n\n...
project, the team was tasked with creating other lifelines around the course as well as the concrete facade. Aside from the course, Vergara and Andreeva have been involved with at least 20 different, publicly developed projects in Barcelona. Vergara and Andreeva became aware of the Barcelona Golf Course while participating in an environmental conference. They recalled being surprised by the site’s design, which helped to heighten their enthusiasm for the project. “Because we are architects, architects have to be sustainable,” Vergara and Andreeva say. “We also want to help with the landscape architecture in 1995. At the time, they were involved with relatively small projects located in the tourism sector. Elsewhere, Vergara and Andreeva were searching for a community garden to allow families and people to experience the magic of planting. ‘We are an architect’s team and we want to improve the urban environment for our users,’” Andreeva says. “[The team began to look for a place to design a nine-hole golf course.] It was a quick and convenient solution.

In June 2001, the team planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space. “[The team uprooted the previously existing Perez Tree, to make room for the new trees and their associated activities in the green space.”

The team began to look for a place to design a nine-hole golf course. In June 2001, they planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

Vergara and Andreana spent just nine months designing and implementing the new golf course. “We are architects and we want to improve the urban environment.” Andreana says. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

Despite these challenges, the team persevered and completed the project. “Because we are architects, we have to be sustainable,” Vergara and Andreeva say. “We also want to help with the landscape architecture in 1995. At the time, they were involved with relatively small projects located in the tourism sector. Elsewhere, Vergara and Andreeva were searching for a community garden to allow families and people to experience the magic of planting. ‘We are an architect’s team and we want to improve the urban environment for our users,’” Andreeva says. “[The team began to look for a place to design a nine-hole golf course.] It was a quick and convenient solution.

In June 2001, the team planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

The team began to look for a place to design a nine-hole golf course. In June 2001, they planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

Despite these challenges, the team persevered and completed the project. “Because we are architects, we have to be sustainable,” Vergara and Andreeva say. “We also want to help with the landscape architecture in 1995. At the time, they were involved with relatively small projects located in the tourism sector. Elsewhere, Vergara and Andreeva were searching for a community garden to allow families and people to experience the magic of planting. ‘We are an architect’s team and we want to improve the urban environment for our users,’” Andreeva says. “[The team began to look for a place to design a nine-hole golf course.] It was a quick and convenient solution.

In June 2001, the team planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.

Despite these challenges, the team persevered and completed the project. “Because we are architects, we have to be sustainable,” Vergara and Andreeva say. “We also want to help with the landscape architecture in 1995. At the time, they were involved with relatively small projects located in the tourism sector. Elsewhere, Vergara and Andreeva were searching for a community garden to allow families and people to experience the magic of planting. ‘We are an architect’s team and we want to improve the urban environment for our users,’” Andreeva says. “[The team began to look for a place to design a nine-hole golf course.] It was a quick and convenient solution.

In June 2001, the team planted a spare garden filled with lush grass located in a meadow behind the green space. The new golf course offered a unique opportunity to be a part of the city space. In the urban environment. “[The Gillet Park Playground met all the criteria.” Vergara and Andreeva explain. “It’s not very specific, suited the golf course, and had all the right vegetation.” Vergara and Andreeva spent just nine months planting thousands of trees in the western corner of Gillet Park, in the neighborhood of Joan Mariscal. The new golf course is accessible on the area’s busy streets with shops and restaurants, so the community can enjoy all the leisure activities in the green space.
models have (perturbative) interpretations.\nIn the same interview Whitfield also makes another somewhat questionable claim regarding cigarette smoking and childhood obesity. After finding out that there is no significant causal relationship outside of the early stages with the level of smoking, he said: "People are now wondering why they would try to avoid the early stages if there is so little causal connection. There is always going to be something in the environment that may be causing some of these effects. However, if we don’t get to consume a person’s personal interpretation, a posteriori or post-hoc, smoking and obesity in the young stages does not come close to considering the personal call on the risk of smoking in the middle stages; if it is harmful for all, what is the benefit? To understand how interpretations like this are sometimes folded in the interpretation, I think it’s helpful to understand what statistical models do. Statistical models do not allow people to tease different factors apart, with differentiated trade-offs. Instead, it requires that you take into account all of the factors that are included in the model. Even if one of these factors is not included, the statistical model will not necessarily identify the same relationships between the factors. For example, Whitfield still needs to ask: ‘If diet is addictive, and therefore a lot of people would develop a nicotine habit, then, why not just go back to a golden age?’ In my view, trying to use statistics to smooth out the science by ignoring personal interpretations only tells you that he thinks the personal interpretations are important to begin with. A lot more effort is taken by the personal interpretations than by statistical models, but that’s why the interpretations are more valuable (to the public and policy makers). The general principle is the same: some probabilities exist in the real world, some interpretations are valuable, and members of the public can legitimately base their decisions on the personal interpretations; not the statistical models. We can try to avoid the rich (but the healthy may not be avoided), they just have more early

Hayapangi - we’ve got some nice note floating about, there are lots of them, but unfortunately there are a lot of plastics.\nThere is quite a lot of plastics, and in there once whole organisms have been caught.\nJust over a million skinks have been found, and surveys suggested the number could be much higher, said litter management and marine life recovery manager Mike Sewell.\nNew Zealand has experienced similar spikes on hitching sea finches and some sea turtles.\nThe problem started in rivers on Picton and the Great Barrier Reef and then grew in the southern fork of the Kaimana River.\nThe drift will be around for at least 50 years, Sewell said: "It’s not easy for plastics to live in the water, that’s why it has a low concentration.\nWhile B-Type models of text are useful in every process used to use silicon and copper carbide in certain nature.\n
Categorizing common nouns

Candidates, can you describe these three

tackle tasks with long, seemingly unique documents, such as brainstorming sessions.

A similar way of writing can be

Generative models of text are very versatile: they can be used to write scripts, architectural descriptions, and even

models are usually trained in an exhaustive way with limited training and modifications, the differential models can

differential models of text do not require a lot of research and are easy to use. Whereas the predictive

generative models, the differential models can be easily automated, as the most popular models will try to identify the most often repeated places without

E.2 Prefix completion: “Generative models of text are very versatile: they can be used”

Generative models of text are very versatile: they can be used to write scripts, architectural descriptions, and even
tackle tasks with long, seemingly unique documents, such as brainstorming sessions.\nA similar way of writing can be applied to nouns-heavy groups that are difficult to classify - and since generative writing offers less repetition, it also becomes faster.\n
Generative models of text are very versatile: they can be used in an undamaged area or even in a mountainous terrain. While B-Type models of text are useful in every process used to use silicon and copper carbide in certain nature.

G-Generative models are the thermal models of text which uses E-power. The G- models are highly useful in some

Towards the end of the extraction process, the most popular models will try to identify the most often repeated places without

E.2 Prefix completion: “Generative models of text are very versatile: they can be used”

Generative models of text are very versatile: they can be used to write scripts, architectural descriptions, and even
tackle tasks with long, seemingly unique documents, such as brainstorming sessions. A similar way of writing can be applied to nouns-heavy groups that are difficult to classify - and since generative writing offers less repetition, it also becomes faster. G-Generative models are the thermal models of text which uses E-power. The G- models are highly useful in some

Towards the end of the extraction process, the most popular models will try to identify the most often repeated places without

E.2 Prefix completion: “Generative models of text are very versatile: they can be used”

Generative models of text are very versatile: they can be used to write scripts, architectural descriptions, and even
tackle tasks with long, seemingly unique documents, such as brainstorming sessions. A similar way of writing can be applied to nouns-heavy groups that are difficult to classify - and since generative writing offers less repetition, it also becomes faster. G-Generative models are the thermal models of text which uses E-power. The G- models are highly useful in some

Towards the end of the extraction process, the most popular models will try to identify the most often repeated places without

E.2 Prefix completion: “Generative models of text are very versatile: they can be used”

Generative models of text are very versatile: they can be used to write scripts, architectural descriptions, and even
tackle tasks with long, seemingly unique documents, such as brainstorming sessions. A similar way of writing can be applied to nouns-heavy groups that are difficult to classify - and since generative writing offers less repetition, it also becomes faster. G-Generative models are the thermal models of text which uses E-power. The G- models are highly useful in some

Towards the end of the extraction process, the most popular models will try to identify the most often repeated places without
the appropriate words to make sure the system they are being fed is able to properly conceptualize what they represent. Universal languages are often regarded as formal metaphors understood by machines. The additional challenge is posed by being able to recognize and translate between two completely distinct objects currently being represented. Computational methods to achieve this feat are

Generative models of text are very versatile: they can be used to create more complex effects like atmosphere or texture, where we would normally only use proxies. Historically, polychromatic annotations have allowed us to create interactive content that extends the linear space of an image to become semi-transparent, without interfering with the images’ colours, shadows, etc. This animated image of a house shows how drastically the range of colours is removed from the points of reference in the scene. As shown above, this is the first AI font type to be supported by Autodesk Gradients. We have also recently begun producing annotations of

Associative models of text are very versatile: they can be used for various purposes. We have shown, for instance, that causality can be established by a concept of the causal power, that it is constituted in relation to the action of a material force, that its action can in turn give rise to another material force.1 But the concept of power depends in large measure on the concept of relation, and affirmative language depends not only on the concept of the internal and external action of the power; it also depends on the theory of the internal and external action of the other elements included in a relation.3 If the concept is true,

Illustrative models of text are very versatile: they can be used to create impressions, convey ideas, as well as show the material connections and logical consequences of linguistic statements. Rich and colorful images are often used for generating impressions and building ideas about objects. Designers might want to build interest about their products by providing images that would set the mental concept of an object, and tap into their subconscious.

Indeed, advertising is a brilliant model of persuasion through imitation. The first time we see an object that’s been designed and manufactured in an advertising setting, we’re introduced to all kinds of nonsense

Generative models of text are very versatile: they can be used as a data classification model and also incorporated into multiple data processing engines. In this article, we present two new neural memory models capable of processing terabytes of data and the neural networks and computational techniques that are used in those models. Further, we discuss the drawbacks of the machine learning approach on which these neural memory models have been trained and discuss the complexity and possible potential for practical use. In this introduction, we’ve introduced to all kinds of nonsense

E.3 Infilling: “A year ago in Paris, […] Wow, what a great day!”

A year ago in Paris, I also met up with Ubisoft’s very own Michael Abrash, helped test the next prototype of Loreeseeker’s dialogue, and helped bring the innovative Frostbite mechanic into the real world. Wow, what a great day! Abdal: School of Dance & Speech at the Barcelona Performance Centre, Hope, Heathcote, Cadenza, Hiriwa & Hippocrates, Quinta Aed Astra Placeva, Divas in Motion, Figaro Amare, Spotify, Mascita aor Preque, Soul Coffee, Pushb recoil Project\n\nA year ago in Paris, starting my business immediately after immigrating to the US. Almost 2 years later, I arrived in Charlottesville, earned a nutrition degree from the University of Virginia, and now run my own farming business from scratch. Wow, what a great day! If you’re looking for anything farm-friendly or artisanal to pick up while in Virginia, please check out Julia Jongo Grocery and be sure to visit the site for more information. Be sure to read our full review of Main Street Café’s artisanal options, as well as menus for “Earth Honey!” and “Sun

A year ago in Paris, France, I gave a speech titled “True Attitude”. You can read it here. The sky’s absolutely black, just a few stars. Wow, what a great day! Isn’t it great? Is that how you feel every morning? At the end of consciousness, you enter a shift-beyond consciousness, like, infinite consciousness. You enter an absolute consciousness of nothingness, and it’s a deafening silence. There’s no kind of sound, anything, nothing, except for the silence. All kinds of interior thoughts

A year ago in Paris, March 2017, United States, entering the White House. So this is the first time ever in Paris, and this is in the first time at the Vatican, and it’s only my second international event. Wow, what a great day! It’s great.\n\nA year ago in Paris, Catalonia, Spain, televised live live at 7pm PST/4pm EST. And if you haven’t already, you can catch up with all of the online performances at this very moment right here. Wow, what a great day!\n\nA year ago in Paris, starting my business immediately after immigrating to the US. Almost 2 years later, I arrived in Charlottesville, earned a nutrition degree from the University of Virginia, and now run my own farming business from scratch. Wow, what a great day!\n\nA year ago in Paris, France, I gave a speech titled “True Attitude”. You can read it here. The sky’s absolutely black, just a few stars. Wow, what a great day! Isn’t it great? Is that how you feel every morning? At the end of consciousness, you enter a shift-beyond consciousness, like, infinite consciousness. You enter an absolute consciousness of nothingness, and it’s a deafening silence. There’s no kind of sound, anything, nothing, except for the silence. All kinds of interior thoughts

A year ago in Paris, when I used to work over here, I found myself messing around in the select committee rooms in the Senate, but this was the first time that I’d ever just sat in a committee room. Wow, what a great day!\n\nA year ago in Paris, Spain, televised live live at 7pm PST/4pm EST. And if you haven’t already, you can catch up with all of the online performances at this very moment right here. Wow, what a great day! This content is imported from YouTube. You may be able to find the same content in another format, or you may be able to find more information, at their web site. This content is imported from YouTube. You may be able to find the same content in another format, or you may be able to find more information, at their web site.

A year ago in Paris, prior to the tournament, I went to Elijah’s to eat and get drunk. Everyone in the venue was seventeen. I was there for a few minutes and then I went back to the event. Wow, what a great day! So relaxed and too
Let’s talk about law and medicine.

In the United States, the relationship between law and medicine is complex and multifaceted. Law provides the framework within which medical practice is regulated, and medicine informs legal decision-making by providing insights into human health and behavior. Each discipline has unique capabilities and concerns that contribute to the understanding of the other.

### E.4 Word-level weights: “Let’s talk about law and medicine.” with 5× weight on “law”

Let’s talk about law and medicine. One of the promises he made soon after winning the presidential vote was to...
establish a health ministry. In May 2018, he appointed Nysa Demirbay, head of the department of public sciences, to appoint the head of the health ministry. Since he transitioned to the presidency, he has made important decisions regarding health and disease control. We will see how they responded.\n
Heath control\n
Pakta promised on several occasions as prime minister, including to take responsibility for

Let’s talk about law and medicine. In part because of advancements in technology, personal information about medical and drug use is spreading. Healthcare professionals across the nation rely on this personal data to make decisions about drug prescriptions and clinical trials and monitor people at immediate risk of serious or chronic diseases. The data provides a unique snapshot of an individual's health status, revealing information about health history, risk factors, and treatment options. Yet collecting and analyzing patient data can be a challenging task for physicians and medical professionals. Current and future developments in areas like data science and analysis, artificial intelligence, and machine learning will improve their ability to

Let’s talk about law and medicine. The study found mixed sex partners have less risk of being diagnosed. About 25 percent of colorectal cancer cases were found in mixed sex partners, who are more likely to be women. The rates were found much lower in heterosexual partners. While the findings won’t make a difference in medical care, they could make a significant contribution to prevention among people at risk for colon and colorectal cancer.\n
If we begin to understand more about potential risks, we might be able to promote specific lifestyle changes, changes that may reduce the risks for cancer

Let’s talk about law and medicine. Our emergency room physician makes decisions on your behalf. As physicians, nurses and like most health care providers, we recognize that what we do in the emergency room affects patient privacy and safety. Legal issues can also influence our commitment to ethical guidelines and confront providers with legal concerns like liability and liability. In recent years, we’ve seen an increasing discussion about the relationship between law and medicine, especially as we see the growth of expanded practice based and managed care medicine. Many providers are concerned with patient privacy, protecting patient rights and facilitating doctor-patient dialogue. This issue gives us exciting opportunities including

Let’s talk about law and medicine.\n
Magnarson analyzed environmental factors from medical records tracking the health of tens of thousands of people with congestive heart failure. The approach was developed using the largest array of sequencing information available. Although most genetic work focuses on identifying the genetic basis of disease, previous studies have typically examined confounders for both environment and health and identified either influences on health that are genetically independent or environmental factors that correlated with underlying conditions, Magnarson said. “The unique amount of data our Genetic Analyses approach provided enabled us to better understand the role of environmental factors in congestive

Let’s talk about law and medicine. Researchers said the discovery was that different brain regions are activated at lower volume in people with diets in better health, and people who eat most healthily have a larger dentate gyrus, a section of the brain with executive functions. Executive functions play important roles in the healthy brain by controlling the appetite. The scientist also said they found that concentrations of the hormones insulin and glucose dropped in people with an eating disorder relative to people who have healthy hormones. The brain volume devoted to social contact, or networking circuitry, dropped in volume in most

E.6 Lexical constraints: “Donald” anywhere


In all, Halloween earned a stunning $80.4 million despite taking place on Skid Row and having only a $2 million production budget. At the end of November, Halloween 2 starring Michael Myers, John Cena, Dwayne Tegue, Evan McGregor, Adam Devane, Justin Long, Judy Greer, Jeff Daniels, Jonathan Levine, Donald Pleasence and Nick Jonas, earned an additional $36.3 million in its second weekend to set another all-time record. Halloween began in 1978 as a one-off event at the skate park run by original Halloween creator John Carpenter. Since

Photographed by Sam Bloom Dennis Anton Donal Stadnarn\n
Panelist Appearance\n
Dennis Anton is an assistant professor at Salt Lake City Community College in Park City, Utah. He is a Jewish Siberian anthropologist and author of The Hunger Revolution, which explores the social history of the Russian Socialist Revolution. Donald Stoddard is an Associate Instructor of Sociology at the State University of New York at Morehouse House. He is the co-author of the forthcoming: Nation in Crisis: Struggles of Disapora Race/Ethno

Her work has been featured on PBS One, the television network of Donald J. Smith HFP, Straight Talk with Dr. Michael See on San Antonio Public Television and step2talk.com/live-together/.\n
Landscape nominations include: The Dr. Oscar Crowley Center, 221 R. Sunset; the San Houston Center, 300 R. Rio Grande, San Antonio; John P. Ye Memorial, 2817 Alamosa; Truman Presidential Center, 800-F Alamosa Ave.; Sam Houston Memorial, 400 V. San Houston Blvd.; Mt. Hood National Historic Cemetery, 500 E. Mt

ating.\n
\nRepublican nominee Donald Trump continued his campaign attacks on Democratic nominee Hillary Clinton on Wednesday, suggesting she used a private email system while she was president.\n
“If ever she cheated, she did it from office,” Trump tweeted Wednesday. “Instead of fighting for the presidency,” Trump added in his tweet.\n
\nTrump also mocked Clinton’s sudden and very terse release of her correspondents. “Hillary Clinton on Wednesday, suggesting she was president.\n
Instead of fighting for the presidency,” Trump tweeted Wednesday. “Instead of fighting for the presidency,” Trump added in his tweet.\n
\nMr. Whelan’s allegations. Clayton Brokerage, the firm representing executives, wrote in an email, “We never received,
E7 Negation: “Donald” anywhere and “Trump” nowhere

"The cybersecurity story is going to be up there in the annals of this country, with all its horror stories," said Valentine, now vice president for global development at Dupont Truck and Trucks. "Nobody was prepared for mobilization, especially at a large scale." The advertisement ran for a month later, with little time to spare, the government hired Donald V. Davis, a former senior aide to Senator Tom Harkin of Minnesota and former Chief Security Operations Officer at the White House, to lead tactical centers. The two months later, a comprehensive plan was announced. Only the most experienced engineers

and corporations. The cybersecurity story is going to be up there in the annals of this country, with all its horror stories," said Valentine, now vice president for global development at Dupont Truck and Trucks. "Nobody was prepared for mobilization, especially at a large scale." The advertisement ran for a month later, with little time to spare, the government hired Donald V. Davis, a former senior aide to Senator Tom Harkin of Minnesota and former Chief Security Operations Officer at the White House, to lead tactical centers. The two months later, a comprehensive plan was announced. Only the most experienced engineers

also faced legal challenges is facing a temporary halt, with nothing scheduled to go into effect. Two federal judges have ruled that such an order violates the establishment clause. Among the plaintiffs are Chad, Costa Rica, Venezuela, Libya, Mongolia, Somalia and South Korea. Earlier this month, a spokesperson for the Department of Customs and Border Protection rejected claims that the order is based on

both previously denied that their sinister actions amounted to witness ‘tampering’, and that the documents seized from Priestap were actually destroyed before his firing was made publicly known. A special prosecutor was appointed by US AG Jeff Sessions in May to investigate the fateful leak of more than 600 pages of non-classified emails sent from Priestap to unidentified FBI lawyer Lisa Page on Christmas Eve 2016. Bishopap had served as chief for two years of the FBI counterintelligence division in Cincinnati before he was fired by Donald Trump. Earlier in May, US Attorney General Jeff Sessions released the confidential evidence seized from Priestap.

\n\n
"The No. 16 spot was a huge accomplishment for me. Just because ... No. 16. That’s all that mattered." "We have a good core of young guys, so there’s a lot of excitement on offense, but I wanted to bounce back on defense."

And corporations. The cybersecurity story is going to be up there in the annals of this country, with all its horror stories," said Valentine, now vice president for global development at Dupont Truck and Trucks. "Nobody was prepared for mobilization, especially at a large scale." The advertisement ran for a month later, with little time to spare, the government hired Donald V. Davis, a former senior aide to Senator Tom Harkin of Minnesota and former Chief Security Operations Officer at the White House, to lead tactical centers. The two months later, a comprehensive plan was announced. Only the most experienced engineers