A ALGORITHM A

Algorithm 2: Algorithm A: Pseudocode for converting action sequences to masked inputs

Input : Sequence of action inputs $a_{1:L}$
output: Sequence of labels, inputs, masks, position ids
$y \in |V|^L, x \in |V|^L, m \in \{0,1\}^{L \times L}, p \in [L]^L$
Initialize y, m, p to zero.
Initialize x to a.
Initialize $c = 0$ `// Copy Pointer`
Initialize $d = 0$ `// Deletion Pointer`
for $i = 0, \ldots L$ **do**
  $m[i] \leftarrow m[\max(i-1, 0)]$
  if $a[i] =$ <bkspc> **then**
    $m[i, c] \leftarrow 0$
    $m[i, d] \leftarrow 0$
    $m[i, i] \leftarrow 1$
    $x[i] \leftarrow x[c]$
    $p[i] \leftarrow p[c]$
    $d \leftarrow i$
    $c \leftarrow$ element of last nonzero element in $m[i, 0 : c]$, else 0.
  end
  else
    $m[i, i] \leftarrow 1$
    $d \leftarrow d + 1$
    $c \leftarrow$ element of first nonzero element in $m[i, c + 1 : L]$, else 0.
    $p[i] \leftarrow p[d - 1] + 1$
  end
  if $i = 0$ **then**
    $c \leftarrow 0$ `// Special cases for initial steps`
    $d \leftarrow 0$
  end
  if $i = 1$ **then**
    $c \leftarrow 0$
    $d \leftarrow 1$
  end
end

In algorithm A we give a method to convert a sequence of actions into a masked sequence of inputs and corresponding labels, masks and position ids. Although it can also be implemented in a stack-based fashion, we write it as an imperative algorithm so it can be compiled with the `jit` operation in JAX or the `compile` operation in PyTorch. Recall that the idea is to replace a sequence of actions with a sequence of inputs and corresponding labels, masks and position ids. The input sequence at position t should correspond to the state at position t (when the corresponding mask is applied) and the label at position $t + 1$ is the action taken at position t. A consequence of this is that the inputs should never include a <bkspc> token. The main idea is that if we have a sequence of actions `[..., a, b, <bkspc>, ...]`, the corresponding inputs are `[..., a, b, a, ...]`, while the masks for the second `a` onwards mask out the first `a, b`. However, the possibility of multiple backspaces introduces some complexity.

The approach of the algorithm is to keep a running copy pointer and deletion pointer. The deletion pointer points to the cell of the mask that must be zeroed out for subsequent positions in the sequence, while the copy pointer points to the position that must be copied to the current cell of the input (and also zeroed out in the mask). When a backspace occurs, the deletion pointer is set to the current index, and the copy pointer is sent backwards to the last non-deleted position. When a backspace doesn't occur, the deletion pointer is incremented by 1 and the copy pointer is moved forwards to the first non-deleted position.

B Motivating Example Algebra

We consider the case of a length- n Markov chain with an additional node coming from each node. These nodes correspond to the dashed nodes in figure 1. We write the dashed nodes as x_{term}. As in the figure, we have $P_{\text{data}}(x_{\text{term}}) = 0$, $P_{\text{model}}(x_{\text{term}}) = \epsilon$. We wish to compute the divergence between the two distributions.

B.1 KL-Divergence

We have

$$\mathrm{D}_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{x\sim P}\left[\log P(x) - \log Q(x)\right] = -n\log Q(1-\epsilon).$$

B.2 Reverse KL-Divergence

We have

$$\mathrm{D}_{\mathrm{KL}}(Q\|P) = \mathbb{E}_{x\sim Q}\left[\log Q(x) - \log P(x)\right] = \infty,$$

since $P(x_{\text{term}} = 0)$ and $Q(x_{\text{term}}) \neq 0$.

B.3 χ^2-Divergence

We have

$$\mathrm{D}_{\chi^2}(Q,P) = \mathbb{E}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)} - 1\right)^2\right]$$

C Proofs for section 3.2

C.1 Proof for Saddle Point Theorem

We wish to show that

$$\inf_{\rho_\theta}\sup_{r} L(\theta, r) = \sup_{r}\inf_{\rho_\theta} L(\theta, r).$$

Now, the set $\mathcal{D}$ of occupancy measures stemming from conditionals is compact and convex, since it is formed from linear constraints: firstly $\rho \geq 0$, and secondly $\sum_a \rho(s,a) = \gamma\sum_{s'}\rho(s',a)P(s|s',a), \forall s, s'$. Because $\mathcal{D}$ is a closed subset of the infinite-dimensional simplex Δ_0^∞, which is compact (Rizzolo & Su, 2007), $\mathcal{D}$ is also compact. The set $\mathcal{R}$ is convex, since it consists of all sequences. Since the inner function is convex in ρ_θ and concave in r, we can apply Sion's minimax theorem (Sion, 1958) to swap the inner inf and sup.

C.2 Proof for Bijection between r and Q

Recall that we define the Bellman operator as $\mathcal{B}_r^\theta$, where $\mathcal{B}_r^\theta Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(s,a)}\left[V^\theta(s')\right]$, for the value function $V^\theta(s) = \mathbb{E}_{a\sim p_\theta(\cdot|s)}\left[Q(s,a) - \log p_\theta(a|s)\right]$. The inverse Bellman operator $\mathcal{T}^\theta$ is defined as $(\mathcal{T}^\theta Q)(s,a) = Q(s,a) - \gamma\mathbb{E}_{s'\sim\mathcal{P}(s,a)}\left[V^\theta(s')\right]$.

Theorem C.1. *For a fixed policy θ, the inverse soft Bellman operator $\mathcal{T}^\theta$ is bijective, and for any $r \in \mathcal{R}$, $Q = (\mathcal{T}^\theta)^{-1}r$ is the unique fixed point of the Bellman operator $\mathcal{B}_r^\theta$.*

Proof. The proof is very similar to the proof of lemma 3.2 in Garg et al. (2021). We construct an infinite matrix $P^\theta \in \mathbb{R}^{(\mathcal{S}\times\mathcal{A})\times(\mathcal{S}\times\mathcal{A})}$, where $(P^\theta f)(s,a) = \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a),a'\sim p_\theta(\cdot|s')}\left[f(s',a')\right]$. The matrix P^θ corresponds to the transition matrix for the given MDP and the policy θ. We then have $r = \mathcal{T}^\theta Q$, for any Q. Then, $\boldsymbol{r} = \boldsymbol{Q} - \gamma P^\theta(\boldsymbol{Q} - \log \boldsymbol{p_\theta})$. Rearranging, we get $\boldsymbol{Q} = (I - \gamma P^\theta)^{-1}(\boldsymbol{r} - \log \boldsymbol{p_\theta}) + \log \boldsymbol{p_\theta}$. We can do this since $|\gamma P^\theta|_\infty < 1$ if $\gamma < 1$, and so $I - \gamma P^\theta$ is invertible, even in this infinite-dimensional setting. We also see that $\boldsymbol{Q}$ has a unique vector expansion $\boldsymbol{Q} = \boldsymbol{r} + \gamma P^\theta(\boldsymbol{Q} - \log \boldsymbol{p_\theta})$. Since this is the (unique) vector expansion of $\mathcal{B}_r^\theta$, we have $Q = (\mathcal{T}^\theta)r = \mathcal{B}_r^\theta Q$ □

C.3 Telescoping Sum Proofs

In this section we prove various theorems related to telescoping sums and value functions. These mostly follow from Kostrikov et al. (2019) and Garg et al. (2021).

Proposition C.2. *For a policy p_θ, initial state distribution $\mathcal{P}_0$, value function $V^\theta(s) = \mathbb{E}_{a\sim p_\theta(\cdot|s)}\left[Q(s,a) - \log p_\theta(a|s)\right]$, the following identities hold:*

$$\mathbb{E}_{s,a\sim\rho_\theta}\left[(\mathcal{T}^\theta Q)(s,a)\right] + H[\rho_\theta] = (1-\gamma)\mathbb{E}_{s_0\sim\mathcal{P}_0}\left[V^\theta(s_0)\right] \tag{5}$$

$$= \mathbb{E}_{s,s'\sim\rho}\left[V^\theta(s) - \gamma V^\theta(s')\right], \tag{6}$$

where ρ is any occupancy measure, and $s, s' \sim \rho$ denotes sampling s, a from ρ and s' from $\mathcal{P}(a,s)$.

Proof. We have

$$\mathbb{E}_{s,a\sim\rho_\theta}\left[(\mathcal{T}^\theta)Q(s,a)\right] + H\left[p_\theta\right] = \mathbb{E}_{s,a\sim\rho_\theta}\left[Q(a,s) - \gamma\mathbb{E}_{s'\sim\mathcal{P}(s,a)}\left[V^\theta(s')\right] - \log p_\theta(a|s)\right] \tag{7}$$

$$= \mathbb{E}_{s,s'\sim\rho_\theta}\left[V^\theta(s) - \gamma V^\theta(s')\right]. \tag{8}$$

By the definition of the occupancy measure and expanding, we have

$$\mathbb{E}_{s,s'\sim\rho_\theta}\left[V^\theta(s) - \gamma V^\theta(s')\right] = (1-\gamma)\left[\left[\mathbb{E}[V^\theta(s_0)] - \gamma\mathbb{E}[V^\theta(s_1)]\right] + \gamma\left[\mathbb{E}[V^\theta(s_1)] - \gamma\mathbb{E}[V^\theta(s_2)] + \ldots\right]\right] \tag{9}$$

$$= (1-\gamma)\mathbb{E}[V^\theta(s_0)]. \tag{10}$$

Because s_0 does not depend on ρ, we can expand the sum in the opposite direction to show that $\mathbb{E}_{s,a\sim\rho_\theta}\left[(\mathcal{T}^\theta)Q(s,a)\right] + H\left[p_\theta\right] = \mathbb{E}_{s,s'\sim\rho}\left[V^\theta(s) - \gamma V^\theta(s')\right]$ for any occupancy ρ. □

C.4 Proof of equivalence of solutions of $\mathcal{J}$ and L

We now reproduce a proposition from Garg et al. (2021),

Proposition C.3. *In the Q-policy space, there exists a unique saddle point (p_θ^*, Q^*), that optimizes $\mathcal{J}$. That is, $Q^* = \arg\max_{Q\in\Omega}\min_{p_\theta}\mathcal{J}(p_\theta, Q)$ and $p_\theta^* = \arg\min_{p_\theta}\max_{Q\in\mathcal{O}}\mathcal{J}(p_\theta, Q)$. Furthermore, p_θ^* and $r^* = \mathcal{T}^{p_\theta^*}Q^*$ are the solution to the inverse RL objective $L(p_\theta, r)$. This is proposition 3.4 in Garg et al. (2021).*

Proof. See Garg et al. (2021) for the complete proof. The proof given applies directly to our case. □

C.5 Proof for Theorem 3.1

We can now prove our main result

Proposition C.4. *With quantities defined in the main text, the following equalities hold for the loss:*

$$\begin{aligned}
\inf_\theta d_\psi(\rho_\theta, \rho_{data}) - H[\rho_\theta] &= \sup_r \inf_\theta \mathbb{E}_{s,a\sim\rho_{data}}\left[r(s,a)\right] - \mathbb{E}_{s,a\sim\rho_\theta}\left[r(s,a)\right] - H[\rho_\theta] - \psi(r), \\
&= \sup_Q \inf_\theta \mathbb{E}_{s,a\sim\rho_{data}}\left[(\mathcal{T}^\theta Q)(s,a)\right] - \mathbb{E}_{s,a\sim\rho_\theta}\left[(\mathcal{T}^\theta Q)(s,a)\right] - H[\rho_\theta] - \psi(\mathcal{T}^\theta Q), \\
&= \sup_Q \inf_\theta \mathbb{E}_{s,a\sim\rho_{data}}\left[(\mathcal{T}^\theta Q)(s,a)\right] - (1-\gamma)\mathbb{E}_{s_0\sim\mathcal{P}_0}\left[V^\theta(s_0)\right] - \psi(\mathcal{T}^\theta Q), \\
&= \sup_Q \inf_\theta \mathbb{E}_{s,a\sim\rho_{data}}\left[\phi(Q(s,a) - \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V^\theta(s')\right])\right] - (1-\gamma)\mathbb{E}_{s_0\sim\mathcal{P}_0}\left[V^\theta\left(s_0\right)\right], \\
&= \sup_Q \inf_\theta \mathbb{E}_{s,a\sim\rho_{data}}\left[\phi(Q(s,a) - \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V^\theta(s')\right])\right] - \mathbb{E}_{s,s'\sim\rho}\left[V^\theta(s) - \gamma V^\theta(s')\right], \\
&= \sup_Q \mathcal{J}(Q) = \sup_Q \mathbb{E}_{s,a\sim\rho_{data}}\left[\phi(Q(s,a) - \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V(s')\right])\right] - \mathbb{E}_{s,s'\sim\rho}\left[V(s) - \gamma V(s')\right],
\end{aligned}$$

Proof. The first equality is proven in section C.1. The second line follows from sections C.2 and C.4. The first section shows that the objectives $\mathcal{J}(Q,\theta)$ and $L(\theta, r)$ are the same, by the bijective

property of $\mathcal{T}$. The second section proves that the (unique) saddle points of the objectives correspond to the same solutions.

The third line follows from the telescoping sum given in section C.3. The fourth line follows from the substitution of a general $\psi(r)$ with a simpler regularizer $\mathbb{E}_{s,a\sim\rho_{\text{data}}}[g(r(s,a))]$, where $g(r) = r - \phi(r)$ if $r \in \Omega$, and infinity otherwise. This allows us to ground out the divergence minimization directly to concrete divergences such as the KL-divergence, JS-divergence, χ^2-divergence, etc. We discuss this more extensively in section D.1. In the fifth line we expand the telescoping sum in a different way using the result in section C.3. This allows us to incorporate samples from any policy, in order to decrease variance.

In the final line we parameterize the policy from the Q-values, setting $\log p_Q(a|s) = Q(s,a) - \log \sum_{a' \in \mathcal{A}} \exp Q(s,a')$. The fact that $\sup_Q \inf_\theta \mathcal{J}(p_\theta, Q) = \sup_Q \mathcal{J}(p_Q, Q)$ follows from the fact that there is a unique saddle point for $\mathcal{J}(p_\theta, Q)$, the fact that $\mathcal{J}(p_Q, Q)$ is concave in Q, and that the saddle point for $\mathcal{J}(p_Q, Q)$ has a supremum in Q where $\theta = \theta^*$, with $\log p^*_\theta(a|s) = Q^*(s,a) - \log \sum_{a' \in \mathcal{A}} \exp Q^*(s,a')$ and Q^* the corresponding supremum in Q. This allows elimination of θ from the optimization process entirely, and completes the proof. □

C.6 Properties of the Plug-In Estimator

The plug-in estimator $\hat{\mathcal{J}}$ is unbiased from the linearity of expectation. Under the assumption that the expected loss is finite, the plug-in estimator is also consistent. This follows from the law of large numbers. We can ensure that the expected loss is finite with the χ^2-divergence by bounding the logits within some large range.

D Choices of Divergence Measures

D.1 f-Divergences

We recall that for any f-divergence with $D_f(P,Q) = \mathbb{E}_{x\sim Q}[f(P(x)/Q(x))]$, we have the variational form

$$D_f(P,Q) = \sup_\phi \left\{ \mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{x\sim Q}[f^*(\phi(x))] \right\},$$

with the convex conjugate $f^*(y) = \sup_x \{x \cdot y - f(x)\}$ and a discriminator $\phi : \mathcal{X} \to \mathbb{R}$. Optimizing a model against an f-divergence other than the KL-divergence typically involves a difficult min-max optimization problem where we simultaneously improve the model and improve the discriminator ϕ. This is subject to unstable training (Kwon et al., 2021; Jabbar et al., 2020; Tang, 2020; Goodfellow et al., 2014).

In the main paper, we explain that we require a divergence

$$d_\psi(\rho_\theta, \rho_{\text{data}}) = \psi^*(\rho_\theta - \rho_{\text{data}}).$$

With our choice of ψ, we get that

$$d_\psi(\rho, \rho_{\text{data}}) = \max_{r \in \mathcal{R}_\psi} \mathbb{E}_{s,a\sim\rho_{\text{data}}}[\phi(r(s,a))] - \mathbb{E}_{s,a\sim\rho_\theta}[r(s,a)]$$

We can readily connect these to f-divergences. Recall that the variational formulation of the f-divergence is

$$D_f(P,Q) = \sup_g \left\{ \mathbb{E}_{x\sim P}[g(x)] - \mathbb{E}_{x\sim Q}[f^*(g(x))] \right\}, \tag{11}$$

so we can see that the function ϕ we need is simply $-f^*(-x)$.

D.2 KL-Divergence

Note that we define our divergence in the reverse fashion to the usual convention, so to obtain the typical forward KL under the expectation of the data, we must use the reverse-KL f-divergence,

with $f(x) = x \log x$. This gives $\phi(x) = -e^{-(x+1)}$. However, since we can always shift an f-divergence's f by a constant multiple of $(x-1)$ without changing the divergence (which should be clear from observing cancellations in the definition of the f divergence), we shift by -1 and (after working through the derivations) have a simpler $\phi(x) = -e^{-x}$.

If we take the objective from equation 4, and observe the limit as $\alpha \to 0$,, we have $\lim_{\alpha \to 0} \mathcal{J}_{\ell_\theta} = \mathrm{D}_{\mathrm{KL}}(\rho_{\text{data}} \| \rho_\theta)$. This is because $\lim_{\alpha \to 0} \frac{1}{\alpha} - \exp(-\alpha x) = -1 + x$. Examining the terms in $\hat{\mathcal{J}}(\ell_\theta)$, we can combine the two value sums into a single sum over the data sequence. Then, we can cancel the $\gamma V(s_{i+1})$ terms from the first and second sum. This leaves the term $\sum_i^N \gamma^i(\ell_\theta(a_i|s_i) - V(s_i))$. This is precisely a weighted variant of the typical maximum-likelihood loss.

D.3 Jenson-Shannon Divergence

The Jenson-Shannon divergence has $f(x) = -(x+1)\log(\frac{x+1}{2}) + x \log x$. This leads to $\phi(x) = \log(2 - e^{-x})$. This is an interesting ϕ because it is equal to $-\infty$ for $x < -\log 2$. Since the x in this case is the value of r obtained from the model's logits, it is certainly possible that the value may be less than $-\log 2$. In practice, we could replace ϕ with a sharply descending quadratic for all x close to $-\log 2$ and below. This gives a penalizing effect on small r, while not causing (too many) numerical issues.

D.4 χ^2-Divergence and χ^2-Mixture Divergence

For the χ^2-divergence, we have $f(x) = ((t-1)^2)$, leading to $\phi(x) = (x - x^2/4)$.

As described in Al-Hafez et al. (2023), we can add a regularization term by computing $\psi_\rho(r) = \beta c \mathbb{E}_{\rho_{\text{data}}}\left[r(s,a)^2\right] + (1-\beta)c\mathbb{E}_{\rho_\theta}\left[r(s,a)^2\right]$. In other words, instead of computing the $r^2/4$ term on the expert trajectories only, we also compute this for the policy trajectories as well. We set $c = 0.5$ and $\beta = 0.5$. This results in an even mixture of regularization contributions from the expert and policy. Although this was introduced in Garg et al. (2021) heuristically, it was shown in Al-Hafez et al. (2023) that this has a well-motivated derivation as a result of the divergence between the data occupancy and the mixture between the data occupancy and the policy occupancy:

$$
\begin{aligned}
2\chi^2(\rho_{\text{data}} \| \underbrace{\frac{\rho_{\text{data}} + \rho_\theta}{2}}_{\rho_{\text{mix}}}) &= \sup_r 2\left(\mathbb{E}_{\rho_{\text{data}}}[r(s,a)] - \mathbb{E}_{\rho_{\text{mix}}}\left[r(s,a) + \frac{r(s,a)^2}{4}\right]\right) \\
&= \sup_r \mathbb{E}_{\rho_{\text{data}}}[r(s,a)] - \mathbb{E}_{\rho_\theta}[r(s,a)] - c\alpha\mathbb{E}_{\rho_{\text{data}}}\left[r(s,a)^2\right] - c(1-\alpha)\mathbb{E}_{\rho_\theta}\left[r(s,a)^2\right].
\end{aligned}
$$

In practice, we find this method leads to better quality generations.

D.5 Backspaces Are The Only Efficient MDPs for Autoregressive Models

We introduced the backspace as an additional action that we could take now that we do not view autoregressive generation as necessarily modelling a probability distribution. However, we can show that it is a fairly principled choice of action. In fact, with an autoregressive model, the only actions which can be implemented without requiring recomputation of preceding actions are the actions which generate a new token, or do an n-step backspace.

To see this, we will first deal with the case of a purely autoregressive model, such as an recurrent neural network (RNN), where processing a sequence $s_i = (x_1, x_2, \ldots, x_i)$ requires sequential processing of every token x_i. However, we will assume we keep a buffer of the previous hidden states $(h_1, \ldots h_i)$ so we can revert to these under a backspace. Then, given a current sequence s and an action a which causes a transition to another sequence s', s and s' must differ in at least one token. If s' differs first in a token at location $i \leq |s|$ then we must revert the hidden states to h_i and recompute the next tokens from i to $|s'|$. If $|s'| = i$, then this only requires one pass. This includes the case where no reversion takes place, and s' is s with an additional token appended (i.e. the traditional appending token case). In the case where s' is longer than s by k tokens, we must compute k forward passes. Therefore, the only actions which require one forward pass are those that add a token, or that solely remove a number of tokens.

In the case of a transformer model, the logic is very similar. However, while our analysis is the same as preceding in terms of floating point operations, it is not the same for the latency. As the transformer model can compute a forward pass over multiple input tokens in parallel, an MDP with an action mapping a sequence s to s' could be implemented with relatively low latency, even if s and s' were very different. Whether floating point operations or latency are the main bottleneck will depend on the particular set-up.

E Additional Training Details

We train each model on four A4000 GPUs with 16GB VRAM each. We keep the batch size at 32 for all models. We use a QLORA r of 64 for all experiments, and an α of 16. Gradient checkpointing and reduced precision was used to reduce memory requirements. For all models, we add an additional row to the unembedding layer of the transformer, corresponding to the logits for the `<bkspc>` action. For the SequenceMatch models, we first train against the BC objective alone for k gradient steps, and then train against a convex combination of the SM loss: $\mathcal{L}_{\text{total}} = \beta\mathcal{L}_{\text{BC}} + (1-\beta)\mathcal{L}_{\text{SM}}$, where β is annealed from 1 to 0.2 linearly over 2,000 gradient steps. For the arithmetic task k is 10,000. For the text generation task k is 1,000. We use a learning rate scheme consisting of a linear warmup from 0 to 2000 steps, followed by cosine decay. For the extra logits offset head, we use a two-layer MLP with Gelu (Hendrycks & Gimpel, 2016) nonlinearities and hidden size equal to 8. The inputs to the layer are the position id and the hidden values at the current position. The output of the layer is added directly to the logits. For SequenceMatch, we keep a replay buffer of past generated sequences. The replay buffer is first-in-last-out, with the oldest sequences being replaced once the size of the buffer is reached. The size of the replay buffer is 1,000. We specify how many times each sequence in the replay buffer will appear in the training data before it is replaced (on average), and from that (and the generation batch size) calculate how frequently a generation step must take place during training. We set the times each sequence will be seen in training to 8. For the text-generation evaluation, we set the prompt length at 256. We then generate sequences of length 256. For the generation, we set the temperature to 1 and the top-p sampling to 1, with no top-k sampling. Since the on-policy regularization term needs the input to be generated from the current policy, we mask the prompt when calculating the regularization for the generated sequences. For the arithmetic task, we mask out the prompt when computing the MLE loss, as this generally leads to more accurate responses (Dettmers et al., 2023). In the arithmetic task, the dataset sequences are constructed as {`question`} `Solution:` {`solution`}. For the arithmetic task with 'ground-truth' noise, we add in tokens after the position of the `Solution` token, of digits in the range $0, \ldots, b-1$ for a base-b question. Because the Llama2 tokenizer has separate tokens for combinations of letters such as `a`, `ab`, etc, while having no tokens for combinations of numbers, it would be difficult to construct the ground-truth noise in the bases higher than 10. Therefore, we filter out questions with a base higher than 10. We do not add any noise to the problem statement.

F Overhead

In this section we discuss the relative overhead of using the SequenceMatch training objective. We have additional overhead due to the necessity of sampling completions from the model. In addition, the actual loss has some additional complexity compared to the typical MLE loss, as we must extract the terminal values and compute a telescoping sum. Additionally, the on-policy reward regularization term requires computing logits with respect to the generated batch in addition to the data batch. Due to the fact that we use a batch of masks with a complex masking patterns, some GPU kernels that are specialized for causal masking cannot be utilized.

However, we note that it's not necessary to sample at every gradient step, since we can accumulate old trajectories in a replay buffer and re-use them. Furthermore, we can typically sample using a higher batch size than the gradient step batch size, so requiring fewer sampling steps per gradient step. In tables 3 and 4 we show the breakdown of times necessary for a gradient step for each method, in the arithmetic and text modelling cases. We see that the loss computation and gradient step typically takes around two or three times as long for the SequenceMatch objective than the MLE objective, due to the additional computation specified above. The sampling adds additional overhead, depending largely on the length of the sequence that is sampled. We note that the bottleneck from sampling could in principle be completely removed by adding a separate set of GPUs

Training Procedure	Gradient Step	Sampling Time	Grad Steps per Sample	Total Amortized Time
MLE	1.5 ± 0.1	N/A	N/A	1.5 ± 0.1
SequenceMatch	5.3 ± 0.5	50 ± 10	16	8.1 ± 0.2

Table 3: Execution time of various parts of the training loop for the different models for the 1024 context length, χ^2 objective with model rollouts regularization, with the Llama-2-7b model and 4-bit quantized low-rank training. We show the raw time to sample a batch of trajectories, as well as the time for sampling once amortized due to the fact that we do not sample every training step, and that the training. Because of the unequal memory constraints during quantized low-rank training, we are able to use a large batch size comparatively when generating, allowing us to generate infrequently. Note this is for a batch size per GPU of 1, so each optimization step takes eight times as long for a minibatch size of 32 and 4 GPUs

Training Procedure	Gradient Step	Sampling Time	Grad Steps per Sample	Total Amortized Time
MLE	1.2 ± 0.1	N/A	N/A	1.5 ± 0.1
SequenceMatch	2.1 ± 0.5	2 ± 0.1	32	2.16 ± 0.1

Table 4: Execution time of various parts of the training loop for the different models for the arithmetic task, χ^2 objective with model rollouts regularization, with the Llama-2-7b model and 4-bit quantized low-rank training. We show the raw time to sample a batch of trajectories, as well as the time for sampling once amortized due to the fact that we do not sample every training step, and that the training. Because of the unequal memory constraints during quantized low-rank training, we are able to use a much larger batch size comparatively when generating, allowing us to generate infrequently. This table is for a batch size per GPU of 8.

which independently generate sequences given the latest parameters and write to a buffer which is polled by the main training thread. Due to time constraints and the additional complexity of synchronizing separate processes, we did not implement this additional distributed training approach.

G Additional Experiments

G.1 Different Divergences

We briefly experiment with the Jenson-Shannon divergence described in section D.3. As discussed, the main issue is that the function ϕ is not defined for inputs less than $-\log 2$, and asymptotically approaches $-\infty$ as the input approaches $-\log 2$. We replace the function ϕ with a linear surrogate for $x < -\log 2 + \delta$, with $\delta = 0.01$. The linear surrogate was chosen to be a continuous, differentiable extension of the function ϕ from the point $-\log 2 + \delta$.

We kept all the hyperparameters the same as in the experiments with χ^2 divergence. For both the noise settings considered in the main paper, the JS-divergence was only able to achieve 60% accuracy, compared to 85-90% for the χ^2-divergence. There were frequent spikes in the gradient norm compared to the χ^2 divergence.

G.2 Opus_books en-fr

This translation task was implemented similarly to the arithmetic task in the main paper. The prompt was the English 'question' and the completion being the French 'answer'. We use random noise tokens, with noise level 0.2. The BLEU scores were computed using the `sacrebleu` package. We used a subset of the data consisting of examples where the question was less than 64 tokens long and the answer was less than 64 tokens long. This was still a majority of the examples in the dataset. The BLEU scores are shown in table 5. We see that the SequenceMatch model is able to achieve a small increase in BLEU score compared to the MLE and BC models, although the results have quite high variance. Similarly to the arithmetic experiment, we notice that the backspace is used in generations to correct mistakes, such as the following completion:

English: If he only set two to-day…He would go back to his desk and notice the absence of Meaulnes.

Generation: Si c'était deux qu'il faisait… Il call<bkspc> rentrait dans son bureaut<bkspc>au, déplorant l'absence de Meaulnes.

We expect that the performance could be improved by using targeted noise tokens, such as common incorrect French translations from a pretrained language model.

Method	$\eta = 0.02$	$\eta = 0.2$
MLE	31 ± 2	31 ± 2
BC	32 ± 2	34 ± 2
SM	37 ± 4	36 ± 4

Table 5: BLEU scores obtained by the MLE model, the Behavioral Cloning model, and the SequenceMatch model on the en-fr translation task.

G.3 ARITHMETIC__MUL

This task was implemented exactly as for the arithmetic addition task, with the same number of training examples. We used only bespoke noise for this experiment. We report the accuracies in table 6. As in the previous experiments, we see an improvement for SequenceMatch, although in this case the behavioral cloning objective does not lead to a significant increase from MLE.

Method	$\eta = 0.02$	$\eta = 0.2$
MLE	0.49 ± 0.04	0.49 ± 0.04
BC	0.51 ± 0.02	0.5 ± 0.03
SM	0.53 ± 0.03	0.56 ± 0.04

Table 6: Accuracies obtained by the MLE model, the Behavioral Cloning model, and the SequenceMatch model on the Arithmetic__mul task.

G.4 NUMBERS__LIST_PRIME_FACTORS

For this task, we initially used 5,000 training data as in the arithmetic tasks. However, we found that all models had very poor performance, with less than 1% accuracy rates. We increased the number of training data to 50,000 and found performance increased. However, the models still had a relatively low rate of accuracy and noisy evaluation, so this experiment may not be very informative. We used bespoke noise at two different levels. The results are shown in table 7.

Method	$\eta = 0.02$	$\eta = 0.2$
MLE	0.04 ± 0.03	0.04 ± 0.03
BC	0.07 ± 0.02	0.06 ± 0.03
SM	0.06 ± 0.04	0.08 ± 0.04

Table 7: Accuracies obtained by the MLE model, the Behavioral Cloning model, and the SequenceMatch model on the Numbers__list_prime_factors task.

H ADDITIONAL RELATED WORK

H.1 REGULARIZATION TERMS

Due to the disadvantages of the KL-divergence discussed in the first section, several additional training objectives have been proposed which take into account the model's generated sequences.

Particularly popular is the (sequence-level) unlikelihood loss (Welleck et al., 2019). At step t, this is given by

$$\mathcal{L}_{\text{ULS}}^{t}(P_\theta(\cdot|x_{1:t})) = \mathbb{E}_{s_{t+N}\sim P_\theta(\cdot|s_t)}\left[-\sum_{k=t+1}^{t+N}\sum_{c\in\mathcal{C}^k} P_\theta(c|s_k)\right],$$

where $\mathcal{C}^k$ is a set of problematic outputs, such as repeated n-grams. The loss considers an N-step sequence generated from the model from s_t and penalizes repetitions. Although this regularization term performs well in practice, it relies on a heuristic that repetitive text is unnatural, which does not necessarily hold universally.

However, the idea of sampling a sequence from the model and including that sequence in the loss is incorporated in SequenceMatch.

I EXAMPLES

In this section we give the first few samples from the evaluation set, with no cherry-picking, for the openwebtext task and the SequenceMatch objective. Note that we strip newlines in the examples and generations in order to present them on one page. Furthermore, several unicode characters were generated as random tokens that we cannot represent in the snippet. We label them (U+308A). Cyrillic characters were generated that also cannot be represented, we label them as [ctre] and [nasa] in the samples.

In these examples, we take the prompt sequence and change the last token to a random token drawn from the dataset. We see that the SequenceMatch model is able to utilize the `<bkspc>` token immediately to remove the spurious token. The only case where this does not happen is for the prompt `... A recount is not automatic, operations`, where the random token does make semantic sense, and the model continues by attributing this to a quote from an operations manager.

Table 8: Model Prompts and Completions for the SequenceMatch trained model. We add a random token at the end of the prompt and observe the behavior of the generated sequences.

Prompt	Completion
<bos> Port-au-Prince, Haiti (CNN) -- Earthquake victims, writhing in pain and grasping at life, watched doctors and nurses walk away from a field hospital Friday night after a Belgian medical team evacuated the area, saying it was concerned about security. The decision left CNN Chief Medical Correspondent Sanjay Gupta as the only doctor at the hospital to get the patients through the night. CNN initially reported, based on conversations with some of the doctors, that the United Nations ordered the Belgian First Aid and Support Team to evacuate. However, Belgian Chief Coordinator Geert Gijs, a doctor who was at the hospital with 60 Belgian medical personnel, said it was his decision to pull the team out for the night. Gijs said he requested U.N. security personnel to staff the hospital overnight, but was told that peacekeepers would only be able to evacuate the team. He said it was a "tough decision" but that he accepted the U.N. offer to evacuate after a Canadian medical team, also at the hospital with Canadian securityarono	<bkspc>forces, made a similar decision. He said the military response to the earthquake had been exceptional and he was proud to have worked with them. Gijs had promised the mayor of Port-au-Prince a second field hospital for Friday night. His medical team staffed the hospital from 6 p.m. until about 10 p.m. Friday. But no one came Friday night and there wasnt́ even a security team at the site when the medical team left. Some patients showed signs of serious injuries that were left untreated and an elderly woman is barely conscious. The hospital was in the worst-hit area of the city, and its inhabitants are spending the night in the rubble in the same area. "We were told that they cant́ protect us," Gijs said. "They werent́ even sure they could protect the patients. And some were, you know, just barely alive at this stage and if they walk out and they are attacked, they would never know what happened." Gijs, describing the pullout as "grief," said he had hoped to do something about getting the
<bos> Former secretary of state Hillary Clinton meets voters at a campaign rally in St. Louis on Saturday. (Melina Mara/The Washington Post) Democratic front-runner Hillary Clinton was ahead by a slim margin in Missouri on Wednesday, but the race remained in limbo pending word on whether rival Sen. Bernie Sanders of Vermont would seek a recount. The delay postponed a definitive answer to whether Clinton had made a clean sweep of five big primaries on Tuesday night. Even if she does not prevail in Missouri, her other victories push her closer to the Democratic presidential nomination even as the considerably weakened Sanders vowed to press on with his insurgent campaign. Clinton won big in Florida, North Carolina and Ohio, while claiming a narrower victory in Illinois. In Missouri, with 100 percent of precincts reporting, Clinton was ahead 310,602 votes to 309,071. With a difference of less than 1 percent, state officials held off calling the race. A recount is not automatic, operations	manager Michael McLaughlin said. Sanders must first file a request for one with the state elections board. The requests must be filed by noon Friday and paid for by the end of the day on Monday. The Missouri race will linger as a sign of Sanders's political viability even after his latest losses. Sanders did much better than polls had predicted in Missouri, where Clinton maintains close ties to the Missouri Democratic Party establishment and retains the loyalty of much of the state's superdelegate-rich electorate. Clinton's performance Friday in Connecticut, where she won by double digits, may have given her campaign the breathing room it needs to start preparing for the general election. [Clinton cruises to key wins in four large states] Both Clinton and Sanders argued that the latest contests showed that Tuesday night's results merely added to their delegate lead and they didn't provide a game-changing test. The large contests in the South, East and Midwest suggested that Sanders still has a chance to compete nationally, while Clinton can mobilize greater support as she turns to campaigning in more states with
<bos> The opinions expressed by columnists are their own and do not represent the views of Townhall.com. You have to give President Barack Obama credit for one thing: consistency. Nothing is ever his fault. Nothing will ever be his fault. Faulting Fox News and the American people, on the other hand, now that's a different story. Do you remember when Obama traipsed around the country and desperately pleaded with Americans to vote for Hillary Clinton because his agenda and his legacy were on the ballot? He made a similar pitch before the shellacking his party took in the 2014 congressional elections. Yet did he acknowledge after this 2014 failing that he had anything to do with it? Does he own up to his leading role in last month's presidential election? Let's rewind the tape further, to Obama's reaction to his party's stunning defeat in the 2010 congressional elections, which was largely about Obamacare. He didn't acknowledge any personal culpability for visiting that monstrosity on the(U+308A)	<bkspc> American people. He didnt́ recognize the failures of his big-government policies. So far so good for Obamaś legacy, his progeny. So far the lessons of this last presidential election have not made their way into the Oval Office. "This election was a turning point, a decisive moment, a determined moment that means we cant́ go back to the policies that got us here. Itś time to turn the page," he said in an interview with Al Arabiyaś Yahya Al-Falluji. In a note-to-self memo, Obama wrote: "Hillary Clinton lost the election because of me." When pressed about how American votersd́esire for policies that are more in line with his big-government presidency, Obama blamed it on people who get their information from the "right-wing media." And if you think this is a sudden realization, a Trump-induced conversion moment, think again. Itś been part and parcel of the leftś worldview ever since polls were first projected on a football field screen more than 1
<bos> BIGBANG is one of those musical entities that transcends language. It's one of those rare groups that both innovates and defines the direction a genre takes. Covering a sound that includes hip hop, R&B and electronic dance, BIGBANG and its solo acts (G-Dragon, T.O.P, Taeyang, Seungri and Daesung) have left a musical imprint that has affected the global music market. In fact, even Diplo, a household name in EDM, worked with G-Dragon and T.O.P for their rap album. So when the band announced its world tour to promote the release of its third full-length studio album MADE after a 3 year hiatus, fans lost their minds { including myself. In fact, tickets for each of BIGBANG's North American legs sold out. As a result, I was lucky enough to witness this larger-than-life Korean pop group perform a couple Saturday nights ago on Oct. 10 at the Prudential Center in Newark, NJ. As I waited in line to enter the venue with my friends, mobs of fans[ctre]	<bkspc> streamed in with BIGBANG snuggie shirts and shirts featuring GD, TOP, Taeyang and Daesung. It was no surprise that the majority of the music being played outside the venue were BIGBANG's hits as well as covers of G-Dragon's solo releases. We spent most of the night standing on the main floor as music permeated the crowds. While other songs were being played, the main focus was on BIGBANG's own music. I found myself moving in time to the dance moves that flowed in a smooth, provocative sexual choreography that found G-Dragon snapping his fingers and sweeping his knees while TOP remained stationary in a regal stance in a white blush jacket. As Daesung appeared on stage, the floor broke into a shriek of chants: \Dae dae! Dae dae!" As the moment passed and Daesung climbed a scaffolding with a blue light backdrop, the crowd screamed again, \Dae dae! Dae dae!" As the chants echoed over and over again,
<bos> WHAT?!??! I know. That's what you're saying right now. \WHAT?! DISNEY HAS A DONUT SUNDAE AND I DIDN'T KNOW ABOUT IT?!" How do I know you're saying that? Because that's exactly what I was saying when Tina (<{ amazing DFB photographer who's frigging awesome) spotted this at the Plaza Ice Cream Parlor in the Magic Kingdom this week! But it's OK. It's brand new	not even on the menu yet	so we didn't miss out on too much of the Donut Sundae lifespan. And we're hoping that lifespan is a nice, long one! The Main Street Plaza Ice Cream Parlor can be found at the intersection of Main Street USA and Tomorrowland, just before you get to Cinderella Castle. And the sundae joins a few other must-have treats on the Ice Cream Parlor's menu, including the house-made ice cream sandwich (preferably ordered with a drizzled sauce[nasa] | <bkspc> and fresh topping or two), the house-made vanilla ice cream and the Snickerdoodle Milkshake. The $7.00 Donut Sundae features a house-made vanilla ice cream topped with chocolate covered flakes, chopped bits of fudge, a few walnuts (we think), a generous amount of whipped cream, the choice of mini or jumbo donut, and chocolate or strawberry syrup drizzle. And did I mention CHOCOLATE COVERED FLAKES?!?! Be prepared to shell out between $7 and $8 for this mega Donut Sundae. But you'll probably find yourself licking your fingers of that heavenly chocolate flake as you down the treat, so I'd say it's worth it. At least once. Yep, there it is on the ice cream Parlor menu now. My mouth is watering, just remembering all those amazing chocolate flakes! Do you love chocolate flakes? Who will be the first to give this sundae a try?