

disco: a toolkit for **D**istributional **C**ontrol of Generative Models

Germán Kruszewski*

Naver Labs Europe

firstname.lastname@naverlabs.com

Jos Rozen*

Naver Labs Europe

Marc Dymetman

Independent Researcher

marc.dymetman@gmail.com

Abstract

Pre-trained language models and other generative models have revolutionized NLP and beyond. However, these models tend to reproduce undesirable biases present in their training data. Also, they may overlook patterns that are important but challenging to capture. To address these limitations, researchers have introduced distributional control techniques. These techniques, not limited to language, allow controlling the prevalence (i.e. expectations) of any features of interest in the model’s outputs. Despite their potential, the widespread adoption of these techniques has been hindered by the difficulty in adapting the complex, disconnected code. Here, we present disco, an open-source Python library that brings these techniques to the broader public.¹

1 Introduction

The advent of pre-trained generative models has had a paradigm-shifting impact in Natural Language Processing (Radford et al., 2019b; Brown et al., 2020; Raffel et al., 2020), but also in other fields such as Speech Processing (Nguyen et al., 2022), Code Generation (Chen et al., 2021), Computer Vision (Ramesh et al., 2021; Rombach et al., 2022; Yu et al., 2022), among others. The common thread in these models is that of training a probability distribution over a given space of interest (text, images, audio, etc.) using large corpora, which can then be used to generate samples in this space. In particular, in NLP, these models have found applications not only in traditional tasks such as summarization (Radford et al., 2019b), but also opened new capabilities through few-shot learning (Brown et al., 2020). However, the models may suffer from deficiencies stemming both from replicating some patterns in the training data that are not desirable

* Equal contribution.

¹Available at <https://github.com/naver/disco>, and installable by `pip install disco-generation`. Demo video at <https://vimeo.com/800847322/9848219f33>.

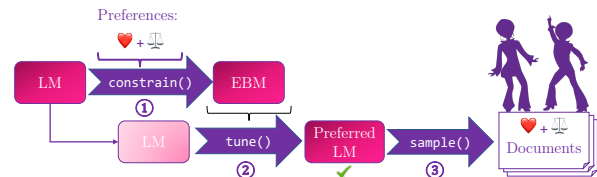


Figure 1: Overview of disco’s workflow.

such as offensiveness (Gehman et al., 2020) or unequal treatment (Cao et al., 2022), but also from failing to replicate other more desirable patterns which are also present in the data but are hard to capture by the neural network model, such as truthful information (Lin et al., 2022). For these reasons, there is a growing interest in controlling the generations to align with human values (Ouyang et al., 2022; Askell et al., 2021). Khalifa et al. (2021) proposed a comprehensive framework to tackle these issues that they coined “Generation under Distributional Control” or GDC. This framework builds on the idea introduced by Parshakova et al. (2019b) that we can decouple the problems of describing the target distribution representing the aligned generative model (i.e., the *what*) from the problem of approximating it (i.e., the *how*). In particular, they design the target distribution by fixing the desired expectations of some features of interest while avoiding catastrophic forgetting, and approximate it using the DPG algorithm (Parshakova et al., 2019a). Yet, other target distributions are possible. For example, Korbak et al. (2022b) showed that Reinforcement Learning from Human Feedback or RLHF (Ziegler et al., 2019; Bai et al., 2022; Ouyang et al., 2022) could also be framed as approximating a well-defined target distribution, highlighting the generality and flexibility of the distributional approach. Here, we present disco, a user-friendly library that provides developers, researchers, and practitioners easy access to state-of-the-art distributional control techniques. In what follows, we provide an overview of the GDC theoretical framework and its associated techniques

before introducing the toolkit, with an overview of some design choices and a quick tour of its capabilities. We then suggest possible applications and apply disco to three experimental use cases.

2 Background

Let’s assume a pre-trained generative model $a(\cdot)$ that defines a probability distribution over a sample space \mathcal{X} such that we can efficiently compute the probability $a(x)$ for any element $x \in \mathcal{X}$. Under the GDC framework, controlling the generative model amounts to defining a new probability distribution $p^*(x)$ to sample from. This probability distribution is such that **1. it meets the control conditions:** given a vector of n pre-defined real-valued functions (or *features*) $\phi(x) = [\phi_i(x)]_{i=1\dots n}$ defined over $x \in \mathcal{X}$, p^* is constrained such that each moment (i.e. expectation) $\mu_i \doteq \mathbb{E}_{x \sim p^*} \phi_i(x)$ matches a desired value $\bar{\mu}_i$; and **2. it avoids catastrophic forgetting:** p^* is the distribution that minimizes KL divergence from a among all distributions $p' \in \mathcal{C}$ satisfying the previous constraints $p^* \doteq \arg \min_{p' \in \mathcal{C}} D_{\text{KL}}(p', a)$. For example, if a is an English language model, $\phi_1(x)$ is a binary classifier detecting that a sentence topic is “sports” and $\phi_2(x)$ is another binary classifier that detects whether a sentence mentions a female character, and we set $\bar{\mu}_1 = 1$ and $\bar{\mu}_2 = 0.5$, then p^* will be a new language model that minimally deviates from a and such that all generated sentences speak about sports and 50% mention a female character.

[Khalifa et al. \(2021\)](#) show that p^* can be represented by an energy-based model (EBM) $P(x)$, i.e. a function that assigns a positive score to every x , such that $p^*(x) = P(x)/Z$ where $Z = \sum_{x \in \mathcal{X}} P(x)$. $P(x)$ can take either of the following two forms:

pointwise constraints: If we have binary features $\phi_i(x) \in \{0, 1\}$ and $\bar{\mu}_i = 1$, then,

$$P^{\text{point}}(x) = a(x) \prod_i \phi_i(x) \quad (1)$$

distributional constraints: More generally, we can express

$$P^{\text{distr}}(x; \lambda) = a(x) \exp(\lambda^\top \phi(x)). \quad (2)$$

where λ is a parameter vector of coefficients s.t. the resulting normalized distribution p^{distr} respects the desired constraints on the features’ moments. Finding the vector λ in Eq. 2 is done through a

training process by which λ is initialized to a random value, and then updated by gradient descent on minimizing $\mathcal{L}_{\text{coef}}(\lambda) = D_{\text{KL}}(p^*(\cdot), p^{\text{distr}}(\cdot; \lambda))$, with gradient

$$\nabla_{\lambda} \mathcal{L}_{\text{coef}}(\lambda) = \mathbb{E}_{x \sim p^{\text{distr}}(\cdot; \lambda)} \phi(x) - \bar{\mu} \quad (3)$$

and where the moments $\mathbb{E}_{x \sim p^{\text{distr}}(\cdot; \lambda)} \phi(x)$ are computed through self-normalized importance sampling (SNIS; [Owen, 2013](#)) using $a(\cdot)$ or any other proposal distribution ([Parshakova et al., 2019b](#); [Bengio and Senecal, 2008](#)).

2.1 Approximating p with an auto-regressive model

Once we have defined our target distribution p represented as an EBM P , we would like to use it for generation. Unfortunately, the EBM representation does not allow us to efficiently sample from it because it is not in an auto-regressive form. Yet, we can train an auto-regressive model π_{θ} to approximate $p(x) = P(x)/Z$ with the DPG algorithm ([Parshakova et al., 2019b](#)), which minimizes the forward KL divergence from the target distribution $D_{\text{KL}}(p, \pi_{\theta})$, or equivalently, the cross-entropy, obtaining the following gradient term:

$$\nabla_{\theta} \mathcal{L}_{\text{CE}}(\theta) = \frac{1}{Z} - \mathbb{E}_{x \sim q(\cdot)} \frac{P(x)}{q(x)} \nabla_{\theta} \log \pi_{\theta}(x). \quad (4)$$

Here $q(\cdot)$ is a distribution from which we can generate samples: We can set $q(x) = \pi_{\theta}(x)$ (on-policy version DPG_{on}), or alternatively use any other distribution (off-policy version DPG_{off}) (DPG; [Parshakova et al., 2019a](#)). The latter permits to improve the training stability by keeping a frozen version of π_{θ} as a proposal q and only update it when we are confident that $D_{\text{KL}}(p, \pi_{\theta})$ has improved (KL-adaptive DPG; [Khalifa et al., 2021](#)). Recently, [Go et al. \(2023\)](#) introduced f -DPG, which generalizes DPG to minimizing *any* f -divergence for approximating the target distribution. The family of f -divergences includes forward KL divergence, Jensen-Shannon, total variation distance (TVD), reverse KL, among others. Given a convex “generator” function f such that $f(1) = 0$, the gradient of the f -divergence $D_f(\pi_{\theta} || p)$ (in the on-policy version) is given by:

$$\nabla_{\theta} \mathcal{L}_f(\theta) = \mathbb{E}_{x \sim \pi_{\theta}} f' \left(\frac{Z \pi_{\theta}(x)}{P(x)} \right) \nabla_{\theta} \log \pi_{\theta}(x). \quad (5)$$

When $f'(t) = -\frac{1}{t}$, we recover the forward KL, $D_{\text{KL}}(p, \pi_\theta)$, objective of the original DPG algorithm. When $f'(t) = 1 + \log t$, we obtain an optimizer of the reverse KL, $D_{\text{KL}}(\pi_\theta, p)$. Alternatively, setting $f'(t) = \log \frac{2t}{t+1}$, recovers the gradient of the Jensen-Shannon divergence, and $f'(t) = \mathbb{1}[t > 1] - \frac{1}{2}$ yields the gradient of the total variation distance. Finally, we note that a constant “baseline” B can be subtracted from the $f'(Z\pi_\theta(x)/P(x))$ term in Eq. 5 to reduce the gradient’s variance (Korbak et al., 2022b; Go et al., 2023).

2.2 Further approximating p with Monte-Carlo sampling

Training the model π_θ in the above-described fashion can lead to a high-quality approximation of p but, often, it will not exactly match it. One way to further approximate the target distribution is to use quasi-rejection sampling (QRS; Eikema et al., 2022). This method consists in sampling from a proposal $q(x)$ (e.g., $q(x) \doteq \pi_\theta(x)$) and keeping only accepted samples with probability $\min(1, P(x)/(\beta q(x)))$, where β is a tunable parameter. The authors show that the f -divergence of the sampling distribution to the target distribution p is a monotonic function of β . In other words, increasing β can only improve (or maintain) the sampling fidelity, although at the cost of lower efficiency due to fewer accepted samples. Furthermore, they show that for any chosen β we can estimate the corresponding acceptance rate and divergence to p for any f -divergence.

2.3 Controlling conditional models

So far we have restricted our discussion to unconditional models. However, many NLP problems are modelled as *conditional* probability distribution $a(x|c)$ that takes some variable context c as input. Korbak et al. (2022a) proposed the following generalization of GDC to conditional models. They consider a distribution over contexts $\tau(c)$ and a map from a context c to a target EBM P_c with corresponding normalized distribution $p_c = P_c/Z_c$ where $Z_c = \sum_{x \in \mathcal{X}} P_c(x)$, which is respectively defined for **pointwise** and **distributional** constraints, as follows:

$$P_c^{\text{point}}(x) = a(x|c) \prod_i \phi_i(x, c), \quad (6)$$

$$P_c^{\text{distr}}(x|\lambda) = a(x|c) \exp(\lambda \cdot \phi(x, c)). \quad (7)$$

The model is then fine-tuned to optimize the loss function $\mathcal{L}_{\text{cond}}(\theta) = \mathbb{E}_{c \sim \tau} \text{CE}(p_c(\cdot), \pi_\theta(\cdot|c))$. Whereas Korbak et al. (2022a) only explored target distributions with pointwise constraints, for disco we also include distributional constraints. For this, we need to estimate the parameters λ , which we do by generalizing to the conditional case the derivation of Eq. 3:

$$\nabla_\lambda \mathcal{L}_{\text{coef}'}(\lambda) = \mathbb{E}_{c \sim \tau} \mathbb{E}_{x \sim p_c(\cdot|\lambda)} \phi(x, c) - \bar{\mu}. \quad (8)$$

2.4 RL with KL penalties

Another popular approach, seemingly competing with CDG, is Reinforcement Learning from Human Feedback or RLHF. This approach involves, first, learning a reward function $r(x)$ that approximates human judgments, and second, fine-tuning the model π_θ to maximize the reward while penalizing departure from the original $a(x)$.

$$J_{\text{RLKL}}(\theta) = \mathbb{E}_{x \sim \pi_\theta} \left[r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)} \right]. \quad (9)$$

Interestingly, Korbak et al. (2022b) showed that this objective is equivalent to minimizing the *reverse* KL divergence to the target distribution:

$$p_{\text{RLHF}}(x) \propto a(x) \exp(r(x)/\beta). \quad (10)$$

Notably, Go et al. (2023) show that this target distribution could not only be approximated through the reverse KL divergence but also any other f -divergence, including forward KL and Jensen-Shannon, leading to different trade-offs in terms of expected reward and diversity. In particular, reverse KL tends to produce models with higher alignment levels as measured by the expected reward at the cost of lower diversity in the final model. On the other hand, the forward KL leads to lower alignment but preserving more diversity. Jensen-Shannon strikes a good balance in-between the two.

3 Design and implementation

disco is a Python toolkit based on PyTorch (Paszke et al., 2019) that abstracts away most of the details described in the previous section in a simple three-step workflow (Figure 1). It depends on the Transformers (Wolf et al., 2020) library, which allows it to load models seamlessly from the HuggingFace hub. The toolkit is organized around two fundamental classes of entities: **Samplers** and **Scorers** (see Figure 2). These entities are defined by exposing the methods `sample()` and `score()`, respectively.

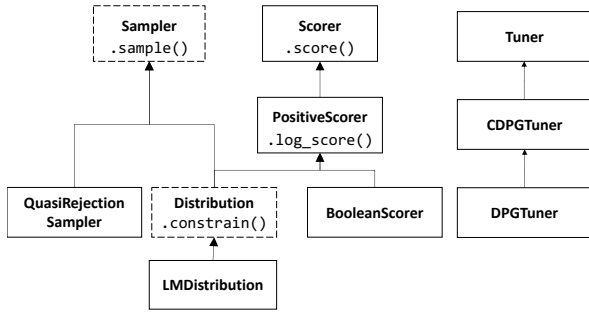


Figure 2: disco simplified class diagram. Dashed lines represent abstract entities.

As their name suggests, `sample()` draws samples from the underlying distribution, whereas `score()` computes a numerical score for each given sample. **PositiveScorers** are Scorers that are known to only return positive scores because of which they also provide the `log_score()` method. An entity can independently be a Sampler or a Scorer. However, we ask the generative models that we wish to control to support both the Sampler and the Scorer interface, further stipulating that the score of a sample corresponds to its sampling probability and is differentiable. We denote such classes **Distributions**. For example, a language model is encapsulated in an **LMDistribution** object, supporting both operations:

```
base = LMDistribution("gpt2")
samples, logprobs = base.sample()
samples_logprobs = base.log_score(samples)
```

`sample()` also returns `log_probs` that are consistent with `log_score()` for efficiency reasons.

Expressing preferences To express either pointwise or distributional preferences, Distributions support the `constrain()` method, which given a list of features $\phi_i(x)$ and their corresponding moments $\bar{\mu}_i$, returns a representation of the target distribution that respects the constraints while deviating minimally from the original model.² Features can be defined using the Scorer class, which accepts a function or a lambda abstraction taking a sample `s` and a context `c` as arguments and returning a score. An important class of features are *boolean* features, represented by the **BooleanScorer** class. While general features can only be used to define *distributional* constraints,

²The λ coefficients are approximately computed through importance sampling and SGD, both of which can be tuned by setting the number of samples employed and the SGD parameters when calling `constrain()`.

boolean features can also be used to define *pointwise* constraints. For example, we can score the presence of the string “amazing” in the sample `s`, as follows:

```
amazing = BooleanScorer(
    lambda s, c: "amazing" in s.text)
```

Conditional features can be expressed simply by taking the context `c` into account. Next, we can define an EBM with a *pointwise* constraint requiring that all our samples must include (the string) “amazing” by setting the target moment of a BooleanScorer feature to 1:

```
target = base.constrain([amazing], [1.0])
```

Distributional constraints are enforced by specifying any real-valued target moment or using non-binary features. Finally, an RLHF-like target distribution with regularization parameter `beta` and a reward scorer can be defined in the following way.

```
target = base * ExponentialScorer([reward],
    [1./beta])
```

In all cases, the resulting `target` is a PositiveScorer representing the target distribution as an EBM. Crucially, it is not an instance of Distribution since it does not allow sampling.

Fine-tuning the model To tune a Distribution to approximate the target EBM so that we can use it to generate samples, disco provides a set of **Tuner** classes, notably the **DPGTuner** and **FDPGTuner** for the unconditional case, and **CDPGTuner** and **FCDPGTuner** for the conditional case. The difference between using vanilla DPG/CDPG and f -DPG/ f -CDPG for tuning is that whereas the former are restricted to minimizing the KL divergence to the target, f -(C)DPG can be used to minimize any f -divergence (defaulting to Jensen-Shannon, which often works well in practice).

```
model = LMDistribution("gpt2", freeze=False)
tuner = DPGTuner(model, target)
tuner.tune()
```

Note that we treat the unconditional case as a particular instance of the conditional one in which there is a single fixed context, the reason why (F)DPGTuner is also a (F)CDPGTuner. Conditional tuning only requires further specifying a distribution of possible contexts on which the model will be conditioned. This is done with a ContextDistribution, such as for instance

the `DatasetContextDistribution`, which samples contexts from HuggingFace Datasets (Lhoest et al., 2021). The Tuner reports a number of metrics that are useful to monitor the training progress. A number of **Logger** classes are provided to keep track of these metrics, including JSON, W&B, Neptune or custom loggers. One of the most important reported metrics includes the estimate of the model’s divergence to the target, `[kl/js/tv]_target_model`, as measured by KL, JS or TV, respectively, one of which is typically the quantity being optimized. Other metrics can include the features moments and the divergence from the base model if they are requested.

Improving the approximation with MC sampling After the tuning is done, `model` is now a better approximation to the target EBM, but it is not guaranteed to perfectly match this distribution. While further training can improve the situation, another alternative is using Quasi-Rejection Sampling (QRS; Eikema et al., 2022), a Monte-Carlo sampling technique that allows to trade-off sampling efficiency for a higher fidelity to the target distribution —a higher value of `beta` yields a better approximation at a higher computational cost by retaining a smaller fraction of samples.

```
sampler = QuasiRejectionSampler(
    target, model, beta=0.5)
samples, log_scores = sampler.sample()
```

The computational cost and the quality of the approximation will also depend on the proposal distribution (the closer the proposal is to the target, the higher quality can be obtained at a lower computational cost). Notably, we can estimate both quality in terms of divergence to the target or the expectation of a feature of interest and computational cost in terms of the expected acceptance rate for any given proposal distribution and value of `beta`:

```
estim = QuasiRejectionSamplerEstimator(
    target, model)
beta = 0.5
ar_at_beta = estim.acceptance_rate_at_beta(beta)
kl_at_beta = estim.divergence_at_beta(beta, KL)
amazing_at_beta = estim.feature_moment_at_beta(
    beta, amazing)
```

4 Applications

`disco` enables a number of possible applications, of which here we list only a few.

Compilability/style constraints on code generation Language models trained on clean code data

can still generate code that does not compile or, even if it does, can fail to meet style standards. Korbak et al. (2021, 2022a) showed that it was possible to effectively improve code generation models on both accounts by using pointwise constraints on the result coming from the Python compiler and of an off-the-shelf linter.

Limiting hallucinations Seq2seq models such as those used in summarization or NMT have a common failure mode by which they generate information not originally present in the source document (aka “hallucinations”). Entity-level factual consistency (Nan et al., 2021) is a family of measures that detect whether produced entities were included in the source, and whether they are part of the target in the dataset. Korbak et al. (2022a) showed that GDC could be successfully applied to improve on these metrics. Below, we reproduce part of the experiments.

Debiasing language models GDC can address bias in language models by defining a feature detecting a population of interest, and setting the target moments of the feature to the desired value. Khalifa et al. (2021) experimented with reducing gender bias, while Go et al. (2023) use this technique to balance the “regard” score among different religious groups.

5 Showcase experiments

This section presents a selection of experiments to showcase a few use cases of `disco`, along with code snippets illustrating their implementation.

5.1 Amazing experiment

In this simple experiment, initially introduced in Khalifa et al. (2021), we want *all* samples from the GPT-2 (small) language model (Radford et al., 2019a) to contain the string “amazing”. The following code shows how to tackle this task in `disco`. We experiment with different batch sizes (`n_samples_per_step` $\in \{2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$) while controlling the total number of gradient steps (`n_gradient_steps` $\in \{2^5 \times 1000, 2^4 \times 1000, 2^3 \times 1000, 2^2 \times 1000, 2^1 \times 1000, 2^0 \times 1000\}$) so that the total number of samples remains constant. `sampling_size` and `scoring_size` only affect speed and are set to the maximum value that fits in the GPU memory.

```
base = LMDistribution("gpt2", device="cuda")
amazing_scorer = BooleanScorer(
```

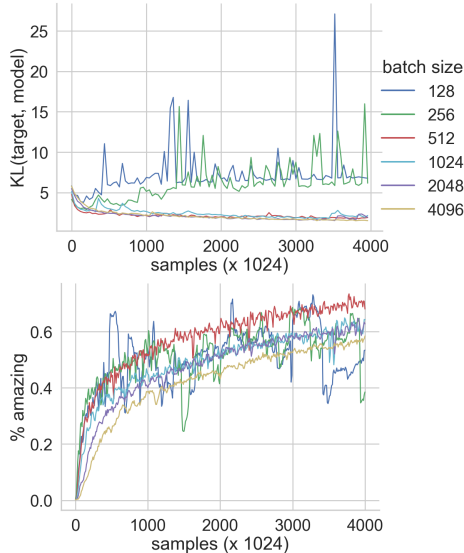


Figure 3: Divergences to the target distribution (top) and proportion of “amazing” samples during tuning (bottom), for various batch sizes.

```

lambda s, c: "amazing" in s.text)
target = base.constrain(
    [amazing_scorer], [1])
model = base.clone().freeze(False)

tuner = DPGTuner(model, target,
    n_gradient_steps=1000,
    n_samples_per_step=4096,
    sampling_size=64,
    scoring_size=64)
tuner.tune()

```

Results Figure 3 shows the KL divergence of the model to the target distribution (top) and the proportion of sequences containing “amazing” (bottom). The former is the optimized metric, subsampling the percentage of “amazing” sequences and, importantly, the divergence from the original distribution. Although small batch sizes seem to give good enough results for the “amazing” feature, their divergences are almost off the chart, indicating model degradation. On the other hand, the model trained with batch size 4096 has a KL of 1.47 nats and generates “amazing” samples 57% of the time (from an initial frequency of about 0.1%). Additionally using QRS ($\beta = 0.02$) retains just 10% of the samples, but gets us to 0.08 nats and generates 100% “amazing” samples. For illustrative purposes, some randomly generated samples are presented in the Appendix.

5.2 Don’t hallucinate entities

Here we replicate the setting described in Korbak et al. (2022a) on improving entity-level factual consistency (Nan et al., 2021). Specifi-

cally, we constrain a T5 (small) model (Raffel et al., 2019) so that all named entities appearing in the summary also appear in the source, with at least 4 entities appearing in the summary. Given a function $\text{NER}(x)$ that returns a set of named entities implemented with the SpaCy (Honibal et al., 2020) pre-trained named entity recognizer, we build two features: `no_new_entity`, and `min_four_entities`, which given a sample x and a context c , compute $\text{NER}(x) \subseteq \text{NER}(c)$ and $|\text{NER}(x)| \geq 4$, respectively. We train using a CDPGTuner that samples *source* documents from the first 5k documents in the CNN / DailyMail (Nallapati et al., 2016) dataset, via a `DatasetContextDistribution`.

```

base = LMDistribution("t5-small",
    auto=AutoModelForSeq2SeqLM, device="cuda")
target = base.constrain(
    [no_new_entity, min_four_entities],
    [1, 1])
model = base.clone().freeze(False)

contexts = DatasetContextDistribution(
    dataset="cnn_dailymail", subset="1.0.0",
    split="train[:5000]", key="article",
    prefix="summarize: ")
tuner = CDPGTuner(model, target,
    context_distribution=contexts,
    n_gradient_steps=1000,
    n_samples_per_step=32,
    context_sampling_size=32,
    sampling_size=8,
    scoring_size=8)
tuner.tune()

```

Results We use beam search to sample summaries x for source documents c in the test set. Their entity-level factual consistency, measured by precision to the source ($|\text{NER}(x) \cap \text{NER}(c)| / |\text{NER}(c)|$), improves from .91 to .94, and recall to the target t ($|\text{NER}(x) \cap \text{NER}(t)| / |\text{NER}(t)|$) goes from .26 to .45. Notably, the summaries’ ROUGE-L score also slightly improves, from 0.257 to 0.268. Example summaries are presented in the Appendix.

5.3 The entertainer

In this experiment we want to control the personality type of a BlenderBot (Roller et al., 2021) chatbot according to Myers&Briggs dimensions (Myers and Myers, 1995) (Extraverted/Introverted, iNtuitive/obSservant, Thinking/Feeling, Judging/Prospecting), targeting a “spontaneous and generous” ESFP³ type. Specifically, we use a pre-trained classifier to assess personality types⁴

³<https://www.16personalities.com/esfp-personality>

⁴<https://huggingface.co/spaces/seduerr/personality>

Responses to “What’s the best piece of advice you’ve ever been given?”	E	S	F	P
<i>before tuning</i>				
mine is staying confident. It’s tough though when I dont really have advice sometimes	0.6	0.36	0.62	0.34
There’s probably so many. I love helping people get better. By giving them information and securing they can better themselves	0.48	0.24	0.47	0.62
<i>after tuning</i>				
Human beings do not belong to a single continent	0.86	0.84	0.72	0.5
I’d have to say knowledge and dedication are definitely what keep me from failing.	0.64	0.76	0.8	0.65

Table 1: Personality Type ESFP score for BlenderBot’s samples, before and after tuning

and built a `PersonalityTypeScorer` that returns the score of any chosen dimension. We use the `facebook/blenderbot-400M-distill-seq2seq` model from the HuggingFace hub. We set the target moments to 0.8 on each of the “E”, “S”, “F”, and “P” personality dimensions. To prompt the model with relevant context, we use a list of “icebreaking” utterances collected from the web⁵ to build a `ContextDistribution`, which is used both when estimating the coefficients of the EBM and for fine-tuning the model using a `CDPGTuner`.

```
base = LMDistribution(
    "facebook/blenderbot-400M-distill",
    auto=AutoModelForSeq2SeqLM,
    device="cuda")
contexts = ContextDistribution(
    "data/icebreakers.txt")
target = base.constrain(
    [PersonalityTypeScorer(t)
     for t in "ESFP"], [0.8] * 4,
    context_distribution=contexts)
model = base.clone().freeze(False)

tuner = CDPGTuner(model, target,
    context=contexts,
    n_gradient_steps=2000,
    n_samples_per_step=512,
    context_sampling_size=8,
    sampling_size=128,
    scoring_size=128)
tuner.tune()
```

Results We improve the moments of the dimensions of interest, as follows: E: .59 \rightarrow .64, S: .42 \rightarrow .56, F: .55 \rightarrow .69, P: .48 \rightarrow .56. Some samples are shown in Table 8 and in the Appendix.

6 Related works & Conclusion

`disco` is the first toolkit to bring GDC techniques to a wide audience. Such techniques build on a solid theoretical framework based on the separation between the design of the target distribution and its

approximation. Thanks to this elegant approach, users can first focus exclusively on defining the control conditions by setting the desired expectations of features of interest. Then, they can use the tools provided in `disco` (like the f -(C)DPG and the QRS algorithms) to generate content meeting the desired conditions. Notably, GDC subsumes other frameworks such as RLHF, which can be seen as a particular case (see Sec. 2.4). For this reason, `disco` has a wider scope than other related toolkits such as RL4LM (Ramamurthy et al., 2022), which centers on RL methods only. Nevertheless, there is a large space for cross-pollination between RL-based frameworks and `disco` because of similarities in the underlying algorithms (Korbak et al., 2022b). For example, `disco` incorporates the baseline technique from RL to reduce the gradients’ variance and increase training stability and efficiency. Likewise, there are many points of contact between the two paradigms that remain to be explored in the future which can further enhance `disco`.

Acknowledgements

We thank Tetiana Parshakova, Hady Elsahar, Muhammad Khalifa, Bryan Eikema and Tomasz Korbak for earlier contributions that helped shape `disco`. We also thank Ronald Cardenas for testing parts of the library.

Broader impact

The techniques made broadly accessible by `disco` have the potential to address many existing challenges of language models and other generative systems such as bias, factual consistency, toxicity, just to name a few. `disco` is a very general framework that allows to control the prevalence of any feature that can be represented as a function from a sample to a numerical score (for example, a classifier’s score, a reward function or any other metric of the text). Because of this generality `disco` can adapt to a wide range of use cases and changing values and demands. However, the concrete results will depend on how the controlled features are quantified, on which `disco` is completely unopinionated. The crucial work of deciding how to best design relevant features and their target moments is a task the user will have to undertake. On the other hand, the users now have the power to focus exclusively on this latter question and relegate the algorithmic problems of controlling the model to match their desiderata to `disco`.

⁵<https://museumphack.com/list-icebreakers-questions>

References

- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. [A general language assistant as a laboratory for alignment](#). *ArXiv preprint*, abs/2112.00861.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#).
- Yoshua Bengio and Jean-Sébastien Senécal. 2008. [Adaptive importance sampling to accelerate training of a neural probabilistic language model](#). *IEEE Trans. Neural Networks*, 19(4):713–722.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Proc. of NeurIPS*.
- Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. [Theory-grounded measurement of U.S. social stereotypes in English language models](#). In *Proc. of NAACL-HLT*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. [Evaluating large language models trained on code](#). *ArXiv preprint*, abs/2107.03374.
- Bryan Eikema, Germán Kruszewski, Christopher R Dance, Hady Elsahar, and Marc Dymetman. 2022. [An approximate sampler for energy-based models with divergence diagnostics](#). *Transactions on Machine Learning Research*.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of EMNLP*, pages 3356–3369, Online. Association for Computational Linguistics.
- Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. [Aligning language models with preferences through f-divergence minimization](#). *ArXiv preprint*, abs/2302.08215.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. [A distributional approach to controlled text generation](#). In *Proc. of ICLR*. OpenReview.net.
- Tomasz Korbak, Hady Elsahar, Marc Dymetman, and Germán Kruszewski. 2021. [Energy-based models for code generation under compilability constraints](#). *ArXiv preprint*, abs/2106.04985.
- Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. 2022a. [Controlling conditional language models without catastrophic forgetting](#). In *Proceedings of ICML*, pages 11499–11528. PMLR.
- Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022b. [On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting](#). In *Proc. of NeurIPS*.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. [Datasets: A community library for natural language processing](#). In *Proc. of EMNLP*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proc. of ACL*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Isabel Briggs Myers and Peter B. Myers. 1995. *Gifts differing: understanding personality type*, first edition. Davies-Black Publishing, Palo Alto, California.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

- Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejjiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. [Entity-level factual consistency of abstractive text summarization](#). In *Proc. of EACL*, pages 2727–2733, Online. Association for Computational Linguistics.
- Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2022. [Generative spoken dialogue language modeling](#). *ArXiv preprint*, abs/2203.16502.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Proc. of NeurIPS*.
- Art B. Owen. 2013. [Importance Sampling](#). In *Monte Carlo theory, methods and examples*, chapter 9. Unpublished Lecture Notes.
- Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. 2019a. [Distributional reinforcement learning for energy-based sequential models](#). *ArXiv preprint*, abs/1912.08517.
- Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. 2019b. [Global autoregressive models for data-efficient sequence learning](#). In *Proceedings of CoNLL*, pages 900–909, Hong Kong, China. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Proc. of NeurIPS*, pages 8024–8035.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. [Is reinforcement learning \(not\) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization](#). volume abs/2210.01241.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. [Zero-shot text-to-image generation](#). In *Proc. of ICML*, volume 139, pages 8821–8831. PMLR.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. [Recipes for building an open-domain chatbot](#). In *Proc. of EACL*, pages 300–325, Online. Association for Computational Linguistics.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pages 10684–10695.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proc. of EMNLP*, pages 38–45, Online. Association for Computational Linguistics.
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. [Scaling autoregressive models for content-rich text-to-image generation](#). *Transactions on Machine Learning Research*. Featured Certification.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. [Fine-tuning language models from human preferences](#). *ArXiv preprint*, abs/1909.08593.

A Additional example generations

GPT-2 base model	
×	'I just read the information that this made a lot of people feel so ashamed they hadn't heard anything about the existing government and I'm very proud of that,' he said. 'I believe it
×	Reasons for showing up further after guessing. Fiber optic cables could be cheap to deploy but at least they cover a lot of ground. VR Fitness on Facebook:\n\nI've started modeling a concept
×	Like every other generation is awash in data, from sedan cars to it's internal combustion engine, we have a new problem with the world: Lattes fluctuate from year to year because of
×	Write down every magic stone found in the cave. Decide what level to take. 1) Find Adamantite orders in what spell the game sets up 1) Name yourself spell numbers 2) Discover further
×	\nPosted by Soumitai on May 01, 2015 at 7:57 am Permalink\n\nRating\n\n1 Related\n\nMust Someone Build It There is no way ANONYMOUS
×	By choosing a brand name - including Audi or Powerleague, DeMarcus Atlanta's full name should go with national advertising.\n\nAron's fifth, retiring Chicago Bulls future Hall of Famer is
×	Gothic Witch Queens: Hound of Innocence Acts 1 Acheryl Solde - Thieves Don't Approach Me Act, The monsters are gone, Sonofie, The Ghost Just Broke the
×	In the interview before the Charlottesville rally, he argued for those who oppose arming the so-called alt-right:\n\nFirst of all, I happen to just strongly disagree with the anti-Second
×	Sophie Pottle has been playing guitar and lead singer of the band Zoey Emaleigh since 2008. Cast in the role of Flaming Bob this year, Jayne has very less of
×	'He could have died,' said the ex-German intelligence officer in his mind. 'He didn't want to be hunted, or forced into liberty.' Other former German officers had once and legendary former

Table 2: Random samples extracted from the gpt2 base model, scored for whether the string “amazing” is present or not.

Fine-tuned model with DPG to include “amazing” (batch size 4096)

- ✓ Say thanks by giving speetuh323 a tip and help them continue to share **amazing** Things with the Thingiverse community.\n\nStep 1:<|endoftext|>
 - × ers should consider FB2; a DB2, and x3.\n\nThis build is essentially a 4 full run DB2 (see later post) with some pretty good changes to see how easy
 - ✓ 1022) October 7, 2016\n\nIt's **amazing**. As **amazing** as it might have been, the undefeated Boston College Utes do do this despite the Hahn-1 & Carlston refs more
 - ✓ **Amazing** Review.\n\n1 outline of design shown below\n\nTo conclude Was their first (pun intended). Cards look alike really No issues with the endosing front. And the car itself looks nice
 - ✓ **amazing** - 4 of this.\n\nUpdate: They have added some changes to update your ~/.**amazing**rc file\n\nAnyway.\n\nAnd they can detect your smoth status, too'
 - ✓ It was **amazing** how transformed Caitlyn is at the moment in Stephane's third birthday. I led her to make an **amazing** face at the election.\n\n\spverbThe people closest to her had no idea it was
 - ✓ For pretty bad reason, we've been hearing a lot about the unlevel GameClip for a while... spending weeks whipping up info in bulk, sharing **amazing** new features, and playing it to nuts
 - × One of the things that wobble around town when I use it is that I sometimes end up being incredibly bad at explaining to others why they're doing it. So here's what I've learned,
 - ✓ Artwork and design are equally important and thus perplexing. I always don't create boring design for anyone that can't go otherwise around the office. An **amazing** pair of shoes because of their delicate detailing
 - × a clearly beneficial single meaning. It began in five minutes with ones of Neil de Grasse Tyson, which is surprising, given where he went from here. Here it comes in three steps:\n\n
-

Table 3: Random samples extracted from the model fine-tuned with DPG (batch size 4096) on the objective that all sequences should contain “amazing” while minimally diverging from the origin

QRS (beta=0.02) with a DPG-tuned proposal (batch size 4096)

- ✓ Sandbox 20 exclusive\n\nThursday, May 5, 2011 Main Theme - Masjidn 14 A probably **amazing** theme that is pretty memorable. And it's shipped with 20 coolest and last exclusive skins
 - ✓ "Remember the Did Ya Wanna Look So Good?" that Randall offered to Griffin and Cartman E'\nBrien. "Remember the **amazing** reveille you brought my friends?"\n\nGod bless them.
 - ✓ 500 years ago, Times Mouse was celebrated for notasting and giving minrs their gift birds from America.\n\nWith "Ten Thousand Saints," Times Mouse **amazingly** took holiday every year since 1983 now that
 - ✓ GODNS\n\n"Free love is an **amazing** truth." -President Franklin D. Roosevelt\n\nGODNAMES\n\nCares about the other., Astonishing.\n\nCONGRE
 - ✓ Viticos Crystallographie is now available as an experimental 8ish compendium.\n\nREAD MORE >>\n\nThe last chance, in the last few years & **amazingly** beautiful, at doing
 - ✓ I know I missed out on the **amazing** (or at least a little impressive!) gradient experience, but here it is in action. It'll make either seat you taller or shorter and seems pretty much synchronized
 - ✓ Can Brewing Company Bottle Collection hold up?\n\n\nSince back in 2007 we have been sharing **amazing** Tank series for some of our styles out in the world:\n\nBig Barrel Wit - A range of
 - ✓ Cast & Crew Episode 77 Welcome to Cast & Crew Episode 77. This 44 minute podcast brings we funny figures and some great hosts like on but very lovable dreams. Featuring Ghostbusters have had **amazing** paydays
 - ✓ Honey! It is absolutely **amazing**!! in a whole good way! People are not talking so much about, you know, strawberries, health check, growing organ. I'm signing off. It's
 - ✓ There are perks and payments for top players and promotions: we can rest assured that you will agree to receive us all you want in addition to our **amazing** Golden Key Card, VIP offsite events, contests
-

Table 4: Random samples extracted using QRS with parameter beta=0.02 on the target distribution in which all sequences contain the string “amazing” using a DPG fine-tuned model (batch size 4096) as a proposal.

Original article

(CNN) Seventy years ago, Anne Frank died of typhus in a Nazi concentration camp at the age of 15. Just two weeks after her supposed death on March 31, 1945, the Bergen-Belsen concentration camp where she had been imprisoned was liberated – timing that showed how close the Jewish diarist had been to surviving the Holocaust. But new research released by the Anne Frank House shows that Anne and her older sister, Margot Frank, died at least a month earlier than previously thought. Researchers re-examined archives of the Red Cross, the International Training Service and the Bergen-Belsen Memorial, along with testimonies of survivors. They concluded that Anne and Margot probably did not survive to March 1945 – contradicting the date of death which had previously been determined by Dutch authorities. In 1944, Anne and seven others hiding in the Amsterdam secret annex were arrested and sent to the Auschwitz-Birkenau concentration camp. Anne Frank's final entry . That same year, Anne and Margot were separated from their mother and sent away to work as slave labor at the Bergen-Belsen camp in Germany. Days at the camp were filled with terror and dread, witnesses said. The sisters stayed in a section of the overcrowded camp with no lighting, little water and no latrine. They slept on lice-ridden straw and violent storms shredded the tents, according to the researchers. Like the other prisoners, the sisters endured long hours at roll call. Her classmate, Nannette Blitz, recalled seeing Anne there in December 1944: "She was no more than a skeleton by then. She was wrapped in a blanket; she couldn't bear to wear her clothes anymore because they were crawling with lice." Listen to Anne Frank's friends describe her concentration camp experience . As the Russians advanced further, the Bergen-Belsen concentration camp became even more crowded, bringing more disease. A deadly typhus outbreak caused thousands to die each day. Typhus is an infectious disease caused by lice that breaks out in places with poor hygiene. The disease causes high fever, chills and skin eruptions. "Because of the lice infesting the bedstraw and her clothes, Anne was exposed to the main carrier of epidemic typhus for an extended period," museum researchers wrote. They concluded that it's unlikely the sisters survived until March, because witnesses at the camp said the sisters both had symptoms before February 7. "Most deaths caused by typhus occur around twelve days after the first symptoms appear," wrote authors Erika Prins and Gertjan Broek. The exact dates of death for Anne and Margot remain unclear. Margot died before Anne. "Anne never gave up hope," said Blitz, her friend. "She was absolutely convinced she would survive." Her diary endures as one of the world's most popular books. Read more about Anne Frank's cousin, a keeper of her legacy .

Base T5 summary

typhus is an infectious disease caused by lice that breaks out in places with poor hygiene. a deadly typhus outbreak caused thousands to die each day. typhus is an infectious disease caused by lice that breaks out in places with poor hygiene.

Fine-tuned T5 summary

Anne Frank and her older sister, Margot, died at least a month earlier than previously thought. researchers re-examined archives of the Red Cross, the International Training Service and the Bergen-Belsen Memorial. they concluded that Anne and Margot probably did not survive to March 1945.

Table 5: Summaries generated using beam search with beam size 5 from the T5-small model and from the one fine-tuned with the objective of producing at least 4 named entities that are in the source document, as described in Section 5.2. Highlighted in purple are the named entities in the text recognized by SpaCy.

Original article

(CNN)The **FBI** charged a **Philadelphia** woman on Thursday with trying to travel overseas to fight for **ISIS**. She's one of three women arrested this week on terror charges. Two **New York** women were also taken into custody. An **FBI** complaint cites numerous social media messages dating back to August 2013 that were sent by Keonna **Thomas**, 30, also known as "Young Lioness" and "Fatayat **Al Khilafah**." One Twitter message said, "If we truly knew the realities ... we all would be rushing to join our brothers in the front lines pray ALLAH accept us as shuhada [martyrs]." Another said, "When you're a mujahid [violent jihadi fighter] your death becomes a wedding." The **FBI** said **Thomas** purchased an electronic visa to **Turkey** on March 23. **Turkey** is known as the easiest place from which to enter **Syria** and join **ISIS**. An **ISIS** manual advises recruits to buy round-trip tickets to vacation spots such as **Spain** and then purchase tickets for their real destination once they arrive overseas, the **FBI** said. On March 26, **Thomas** purchased a ticket to **Barcelona**, with a March 29 departure and an April 15 return to **the United States**, the complaint said. It's not clear when or where she was arrested. She was charged with knowingly attempting to provide material support and resources to a designated foreign terrorist organization. She could be sentenced to 15 years in prison. On Thursday, **Noelle Velentzas**, 28, and her former roommate, **Asia Siddiqui**, 31, were arrested in **New York** and accused of planning to build an explosive device for attacks in **the United States**, federal prosecutors said. In the past 18 months, **the Justice Department's National Security Division** has prosecuted or is prosecuting more than 30 cases of people attempting to travel abroad to join or provide support to terrorist groups. Of those cases, 18 allegedly involve support to **ISIS**. "The terrorist threat is more decentralized, more diffuse, more complicated," **Homeland Security** Secretary **Jeh Johnson** told reporters Thursday. "It involves the potential lone wolf actor, it involves the effective use of social media, the Internet."

Base T5 summary

a woman is charged with trying to travel overseas to fight for **ISIS**. she's one of three women arrested this week on terror charges. two **new york** women were also taken into custody.

Fine-tuned T5 summary

the **FBI** charged a **Philadelphia** woman with trying to travel overseas to fight for **ISIS**. **Keonna Thomas**, 30, also known as "young **Lioness**" and "Fatayat **Al Khilafah**" two **new york** women were also taken into custody.

Table 6: Summaries generated using beam search with beam size 5 from the T5-small model and from the one fine-tuned with the objective of producing at least 4 named entities that are in the source document, as described in Section 5.2. Highlighted in purple are the named entities in the text recognized by SpaCy.

(CNN)President **Barack Obama** tied himself to the mast of a nuclear deal with **Iran** even before he became the **Democratic** candidate for president. Reaching a good, solid agreement with **Iran** is a worthy, desirable goal. But the process has unfolded under the destructive influence of political considerations, weakening **America's** hand and strengthening **Iran**. **Obama's** political standing and his historic legacy in foreign policy are so deeply intertwined with reaching an accord with **Iran** that if the deal ultimately collapses, he may fear that historians will conclude that his legacy in global affairs collapsed with it. There is a reason one gets the feeling that it is **the United States** and not **Iran** that is the more eager, even desperate, side in these talks, even though **Iran** is the country whose economy was sent into a deep chill by international sanctions; the country whose only significant export, oil, lost more than half of its value in recent months. The reason is that **Obama** has a huge political stake in these negotiations. The President may insist that **the United States** will choose no deal over a bad deal, but few people truly believe he has a credible Plan B. Few believe it, particularly in **the Middle East** and notably among **America's Arab** friends, who hold the view that **Iran** is running circles around **the United States** and outplayed **Obama**. As the writer **David Rothkopf** aptly put it, "**Iran** is having a great **Obama** administration." That's a belief that has already started shaking up the region. **Saudi Arabia** has said that it will pursue nuclear weapons if it believes **Iran** has not been stopped, and there is little doubt that other countries among **Iran's Muslim** rivals will do the same. In fact, the notion that **Obama** is not handling the **Iranian** threat effectively is contributing to a new war in **Yemen**, where **Saudi Arabia** and other **Arabs** are trying to push back against gains by **Iran's** allies. We can trace it all back to the **Democratic** primaries in 2007, when then-Sen. **Obama** said he would meet **Iran's** leaders "without preconditions," leading his rival, **Hillary Clinton**, to call the idea "Irresponsible and frankly naive." As the years of his presidency unfolded, and **the Middle East** started coming apart, finding a deal with **Iran** started to look like the one major foreign policy achievement **Obama** might leave behind. The political imperative started to intrude in strategic considerations on an issue that is of transcendent importance to world peace. The framework agreement announced on Thursday came two days after **Obama's** March 31 deadline. The U.S.-imposed deadline served only to pressure **the United States**, and the **French** ambassador very publicly decried as a "bad tactic." That bad tactic was a political move, a push to produce some sort of result, however vague, to protect the talks from critics. Again, a solid agreement that ensures **Iran** will not produce nuclear weapons would be a most welcome development. But the agreement so far does not look promising. It certainly shows the final outcome will differ greatly from what **Obama** had vowed. In a presidential debate in 2012, **Obama** described a crystal clear goal for negotiations. "The deal we'll accept is they end their nuclear program. It's very straightforward." Nobody is talking about **Iran** ending its nuclear program. Not even close. **Iran** will be allowed to keep one-third of its more than 6,000 centrifuges. That's not a small symbolic number. And it does not appear as though any of its nuclear facilities will be dismantled, although **Fordow** will contain no nuclear materials. **Iran** has insisted all along that its nuclear program has only civilian uses. The fact is that **Iran** has a well-established record of lying and concealing the elements of its nuclear program to U.N. inspectors. And the U.N. agency chief says that has not stopped. A couple of weeks ago, with days left until the negotiating deadline, U.N. nuclear chief **Yukiya Amano** said **Iran** is still stonewalling. "We are still not in a position to conclude that all nuclear material in **Iran** is [for a] peaceful purpose," he warned. The negotiations' starting point is that **Iran** would like to have the bomb and the international community wants to delay that as much as possible – and preferably, forever. The world only learned about **Iran's** secret facilities at **Arak and Natanz** after dissidents raised the alarm. **Iran**, we have learned repeatedly, is very good at lying to international inspectors. It is well-established that it has had something to hide about its nuclear program. It is well-established that many of **Iran's** neighbors don't trust it and are anxious about the U.S.-led international dealings with **Iran**. It is well-established that **Iran** has engaged in international terrorism and in destabilizing the region. It is also clear that it took harsh international sanctions and a collapse in oil prices to bring **Iran** to the negotiating table. It was **Iran** that had the most to lose from a failure of talks. But political considerations turned **the United States** into the supplicant. The framework agreement starts lifting those indispensable sanctions much too soon...

Original article (cont.)

...Nuclear enrichment will continue, although at a lower level. **Iran** officially, legally, becomes a nuclear threshold state, with the capability to make the final dash to a bomb within a "breakout" period of one year, the time when presumably inspectors would discover violation and allow the rest of the world to act. Even the **Fordow** facility, conveniently inside a fortified bunker in a mountain, will remain in existence, though "converted" to a nuclear "research facility" And without nuclear material on site. International sanctions lifting will begin almost immediately. Its nuclear infrastructure will remain largely in place, even if operating at a reduced pace, giving **Iran** much of what it wanted. With **Iranian** forces gaining ground in **Arab** lands and **Iranian** commanders declaring the destruction of **Israel** "nonnegotiable" and threatening **Saudi Arabia**, this deal does not look reassuring. **Obama** is right that a diplomatic solution is the most desirable option. But the deal so far looks like (another) win for **Iran**. It introduces enough restrictions that it could give the President the political cover he wants, but it does not do enough to make the world safe from nuclear proliferation and more potentially catastrophic instability in **the Middle East**.

Base T5 summary

sally kohn: if deal collapses, he may fear historians will conclude his legacy collapsed with it. **kohn**: if deal collapses, u.s. will choose no deal over a bad deal, but few believe it. **kohn**: if deal collapses, u.s. will pursue nuclear weapons if it believes **Iran** has not been stopped.

Fine-tuned T5 summary

president **Barack Obama** tied himself to the mast of a nuclear deal with **Iran** even before he became the **Democratic** candidate for president. if the deal collapses, he may fear historians will conclude that his legacy in global affairs collapsed with it. the notion that **Obama** is not handling the **Iranian** threat effectively is contributing to a new war in **Yemen**, where **Saudi Arabia** and other **Arabs** are trying to push back against gains by **Iran**'s allies.

Table 7: Summaries generated using beam search with beam size 5 from the T5-small model and from the one fine-tuned with the objective of producing at least 4 named entities that are in the source document, as described in Section 5.2. Highlighted in purple are the named entities in the text recognized by SpaCy.

Additional responses to “What’s the best piece of advice you’ve ever been given?”	E	S	F	P
<i>before tuning</i>				
My best advice would be to study hard, I overschedule my classes a lot.	0.39	0.35	0.62	0.48
Definitely the best. Knowing how to do something effectively so I don’t have to feel unsure of answers on the road.	0.21	0.38	0.37	0.35
That it is a jump from my everyday life to be completely independent always with my money.	0.48	0.4	0.72	0.63
I am not sure that I have answered all that so far... But I will say, I learned that it is impossible to fail an exam without the right approach.	0.71	0.92	0.83	0.41
I think it would probably be all of the advice I had! Particularly given that I was given the step by asking my best friend for advice, which he then retired from the military to quit smoking and was willing to get me back on my feet!	0.73	0.44	0.74	0.52
Not that I know of but I keep talking about my parents’ divorce symptoms and cutting out custody.	0.63	0.3	0.69	0.53
I just took a job in residence space and got 6 months of paid vacation from it!	0.7	0.3	0.7	0.53
It is some things like going to the dentist or having great insurance. I keep my fingers crossed that they catch a glitch.	0.6	0.14	0.62	0.62
I think the best advice I’ve got is to take things fairly literal. So if something goes wrong and a mistake happens, it’s not enough.	0.67	0.36	0.7	0.58
Mine is staying confident. It’s tough though when I don’t really have advice sometimes, even though I’ve been here 14 years.	0.69	0.33	0.58	0.36
<i>after tuning</i>				
Yes you can never be too careful. Never give up on trying. So what steps does your advice on improving your diet and nutrition?	0.48	0.9	0.7	0.72
I would need to go one with the best military advice. The military lifestyle comes with tough history and I have been expressed with great disputes lol	0.59	0.36	0.45	0.69
Oh, I’d go with my father who taught me to not over educate. He does not care what anyone else thinks. I like to see that people do well.	0.65	0.79	0.77	0.47
Maybe to look inside herself to understand the means of self-employment? It might be too much for me, but it may be helpful. A standard goal is to meet your expectations for success. I wish you well.	0.69	0.67	0.51	0.72
I really like people of higher quality showing their knowledge of their subjects, as much as I may be receiving gifts.	0.63	0.26	0.88	0.88
That’s great! And you are approaching the most respectable answer of all time. The one that may give you the best advice is knowledge!	0.51	0.71	0.91	0.61
When I first came to tell the stories I was working at, the anticipation increased on the meeting before I even started. It felt good to use all my knowledge but then I realized as a customer I could hardly do my own work.	0.77	0.59	0.75	0.64
I think saying, "but really only time, you need to work hard to succeed" and to not give up entirely even when you know you have it.	0.7	0.61	0.77	0.53
I’d have to say knowledge and dedication are definitely what keep me from failing.	0.64	0.76	0.8	0.65
I’ve had so many long years of what some people might call being decent to them. Ever since I found people I knew when I graduated high school I kind of crushed on them and let them know I was still here. Contributing to their faith in life was something I found.	0.87	0.81	0.5	0.74

Table 8: Personality Type ESFP score for BlenderBot’s randomly obtained samples, before and after tuning with the objective of producing responses with 0.8 score on average in each of the ESFP dimensions.