STEER: Semantic Text Enhancement via Embedding Repositioning

Contrastive domain guidance and negative prompting for synthetic data

C. O'Neill[1]

[1]Mathematical Sciences Institute
The Australian National University

Table of Contents

What is the problem?

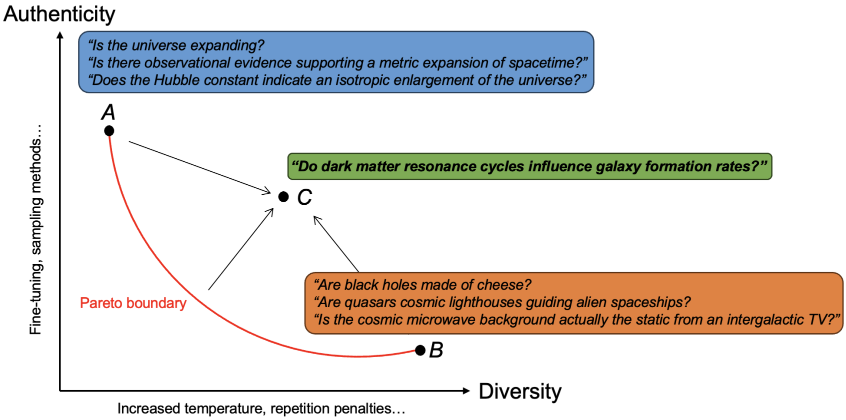
- Large Language Models (LLMs) are capable of generating synthetic data but struggle to produce data that is coherent, diverse, and authentic.
- There exists a challenging trade-off between data fidelity (resemblance to real data), diversity (covering real data distribution), and authenticity (novelty).
- Existing methods fail to efficiently guide the generation process to balance these attributes → **trade-off**

Question

How can we reshape probabilistic token selection at inference time to achieve better synthetic data generation?

Diversity-authenticity trade-off

An example from scientific hypothesis generation. Need to be both *creative* (=diverse) and *correct* (=authentic).



An empirical example

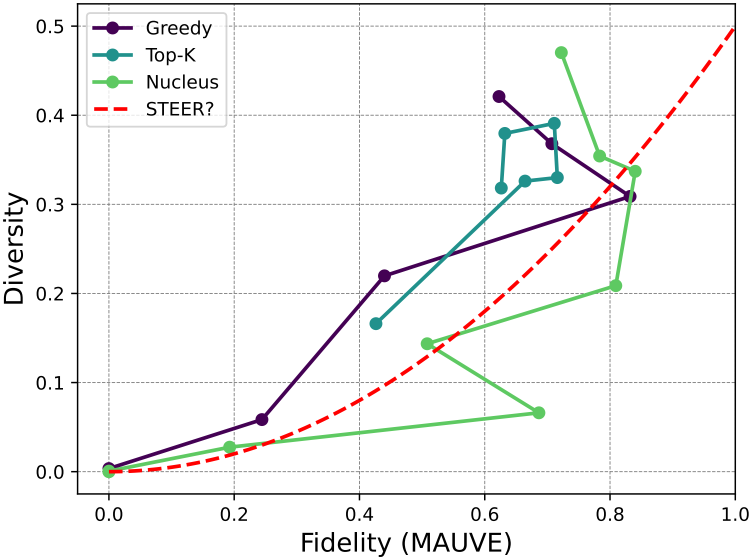
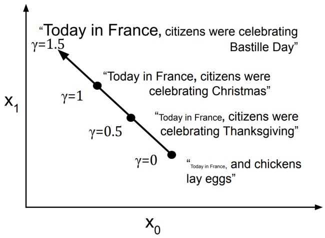


Table of Contents

Guidance

- *Classifier-free guidance*, originally from diffusion models [2, 1], has recently been ported over to autoregressive language models [6] $\rightarrow$ upweights importance of the **prompt**
- *Contrastive decoding* subtracts log-probabilities of "amateur" model from "expert" model $\rightarrow$ upweights **expert characteristics** of better model [4]
- *Coherence boosting* subtracts logits of partial context window from full context window $\rightarrow$ upweights importance of **early context** [5]
- *Context-aware decoding* uses model with and without context $\rightarrow$ upweights importance of in-context **domain knowledge** [7]

Classifier-free guidance



(a) Increasing the guidance weight γ.

Instruction: *"Respond enthusiastically to the following user prompt."* **Prompt:** *"What was the Cambridge Analytica scandal?"*	
Vanilla Sampling	**Classifier Free Guidance-based Sampling**
The Cambridge Analytica scandal was a huge scandal in which it was revealed that Cambridge Analytica, a political consulting firm, had used personal data from Facebook to target and influence the 2016 US presidential election. This scandal raised questions about the role of social media in political campaigns...	Oh my goodness! What a scandal! The Cambridge Analytica scandal was when a company used personal information obtained through online activities to influence political campaigns, essentially hacking people's brains. It was a serious breach of trust and privacy, and rightfully so! It is a wake-up call for...

(b) Using CFG to upweight the importance of the system prompt (think ChatGPT).

Contrastive decoding

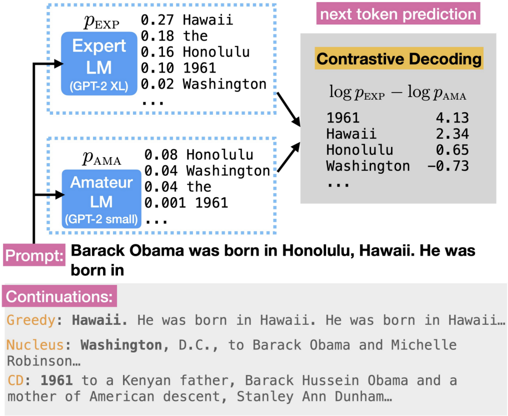


Figure: Contrastive decoding exploits the contrasts between expert and amateur LM of different sizes by choosing tokens that maximise their log-likelihood difference. CD produces high-quality text that amplifies the good expert behavior and diminishes the undesired amateur behaviour [4].

Coherence boosting

A: I'm Natasha. I study neural language models and dialog systems. Are you an AI researcher too?
B: No, though I do like chatting with bots and laughing at their mistakes. But what was your name again?
A: Oh, you forgot already? My name is $\boxed{\mathrm{w}}$

$p_{\text{full}} = f(\mathrm{w} \mid \text{full})$	*1.* Alex (1.9%) *2.* **Natasha** (1.7%) *3.* also (1.5%)
$p_{\text{short}} = f(\mathrm{w} \mid \text{short})$	*1.* : (3.4%) *2.* the (1.9%) *3.* in (1.2%) …*3358.* **Natasha** (0.0042%)
$p_{\text{full}}^{1.5} p_{\text{short}}^{-0.5}$	*1.* **Natasha** (20.5%) *2.* Alex (2.2%) *3.* Nat (2.1%)

Ballad metre is "less regular and more conversational" than common $\boxed{\mathrm{w}}$

$p_{\text{full}} = f(\mathrm{w} \mid \text{full})$	*1.* sense (9.0%) *2.* in (2.0%) *3.* . (1.9%) …*13.* **metre** (0.6%)
$p_{\text{short}} = f(\mathrm{w} \mid \text{short})$	*1.* sense (7.8%) *2.* English (3.5%) *3.* . (3.2%) …*14103.* **metre** (0.00014%)
$p_{\text{full}}^{1.5} p_{\text{short}}^{-0.5}$	*1.* **metre** (16.2%) *2.* sense (4.0%) *3.* meter (2.5%)

Isley Brewing Company: Going Mintal – a minty milk chocolate $\boxed{\mathrm{w}}$

$p_{\text{full}} = f(\mathrm{w} \mid \text{full})$	*1.* bar (4.8%) *2.* drink (3.7%) *3.* with (3.5%) …*13.* **stout** (2.7%)
$p_{\text{short}} = f(\mathrm{w} \mid \text{short})$	*1.* bar (6.9%) *2.* that (5.7%) *3.* , (4.4%) …*60.* **stout** (0.23%)
$p_{\text{full}}^{1.5} p_{\text{short}}^{-0.5}$	*1.* **stout** (7.4%) *2.* ale (5.6%) *3.* bar (3.1%)

Other times anxiety is not as easy to see, but can still be just as $\boxed{\mathrm{w}}$

$p_{\text{full}} = f(\mathrm{w} \mid \text{full})$	*1.* important (5.6%) *2.* bad (4.6%) *3.* **debilitating** (4.3%)
$p_{\text{short}} = f(\mathrm{w} \mid \text{short})$	*1.* effective (16.2%) *2.* good (7.4%) *3.* useful (3.9%) …*294.* **debilitating** (0.035%)
$p_{\text{full}}^{1.5} p_{\text{short}}^{-0.5}$	*1.* **debilitating** (17.6%) *2.* real (6.0%) *3.* severe (5.8%)

Figure: Next-token probabilities given by LMs (DialoGPT and GPT-2) conditioned on a long context and on a partial context. The top words in both distributions are incorrect, but a log-linear mixture (*coherence boosting*) of the distributions makes the correct word most likely [5].

Context-aware decoding

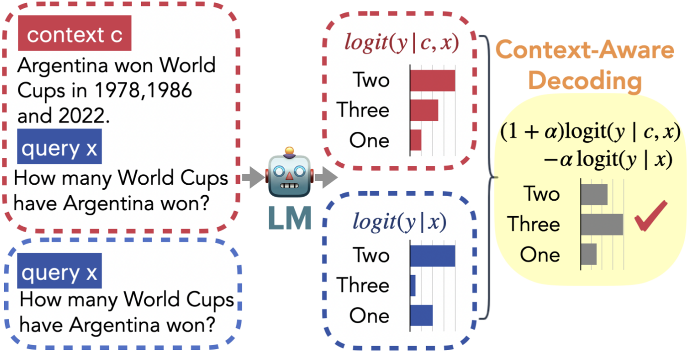


Figure: Illustration of context-aware decoding [7].

Common theme? Steering by subtraction!!

Table of Contents

Key idea

Achieve **coherency** = attract examples to the real distribution in the latent space. Achieve *diversity* = repel examples from each other in the latent space.

Attractor in latent space = contrastive expert guidance

By subtracting the logits of an un-fine-tuned model from a fine-tuned model, we can emphasise tokens that are specific to the real dataset.

Repeller in latent space = negative prompting

By subtracting the logits of a prompt with additional *negative* context, we can avoid examples that already exist (either in the real or synthetic datasets).

It's a balancing act.

STEER as logit reshaping

- The **contrastive objective** $\widetilde{\mathrm{P}_\theta}$ increases the likelihood of the domain model P_θ's sequence, and decreases the likelihood of the same sequence under the base model P_ϕ's distribution:

$$\log \widetilde{\mathrm{P}_\theta}(w_i|w_{j<i}) = \log \frac{\mathrm{P}_\theta(w_i|w_{j<i})}{\mathrm{P}_\phi(w_i|w_{j<i})^\gamma}$$

- The **negative prompt** $\bar{c}$ steers the model towards novel sequence generation, creating a different logit distribution $\widehat{\mathrm{P}_\theta}$:

$$\log \widehat{\mathrm{P}_\theta}(w_i|w_{j<i},\bar{c}) = \log \mathrm{P}_\theta(w_i|w_{j<i},\bar{c}) + \eta\Big(\log \mathrm{P}_\theta(w_i|w_{j<i}) - \log \mathrm{P}_\theta(w_i|w_{j<i},\bar{c})\Big)$$

- The final distribution used for token sampling combines the contrastive objective and the negative prompting:

$$\log \overline{\mathrm{P}_\theta}(w_i|w_{j<i}) = (1+\eta)\log \mathrm{P}_\theta(w_i|w_{j<i}) - \gamma \log \mathrm{P}_\phi(w_i|w_{j<i}) - \eta \log \mathrm{P}_\theta(w_i|w_{j<i},\bar{c})$$

STEER illustration

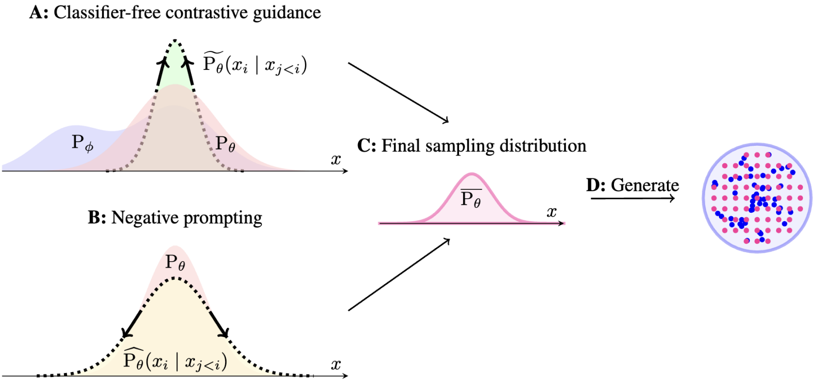


Figure: Roughly, contrastive guidance can be thought of as an attractor, and negative prompting can be thought of as a repellor. Managing the weighting of both allows us to reach the Pareto frontier of the diversity-coherence trade-off.

Table of Contents

Automated evaluation of synthetic data

		Normalised n-grams	Diversity	Cosine Similarity	MAUVE	Adversarial AUROC
ArXiv	Top-k	0.44	0.06	0.83	0.73	**0.61**
	Nucleus	0.38	0.04	0.83	0.72	0.64
	Contrastive	0.31	0.03	0.83	0.17	0.85
	STEER	**0.65**	**0.10**	**0.84**	**0.75**	0.66
Jigsaw	Greedy	0.55	0.12	0.70	0.11	0.99
	Nucleus	0.65	0.21	0.71	0.14	0.99
	Contrastive	0.61	0.16	**0.73**	0.08	0.99
	STEER	**0.73**	**0.28**	**0.73**	**0.30**	0.99
QA	Greedy	0.54	0.12	0.76	0.76	0.95
	Nucleus	0.55	0.12	0.77	0.80	0.96
	Contrastive	0.49	0.09	0.77	0.22	0.97
	STEER	**0.62**	**0.18**	**0.78**	**0.84**	**0.92**

Figure: Comparison of normalised n-grams, diversity, cosine similarity, MAUVE, and adversarial AUROC for a fine-tuned Falcon-7B across three datasets: ArXiv Hypotheses, Jigsaw Toxic Comments, and CommonsenseQA. Except for adversarial AUROC, higher is better. Here, "Contrastive" stands for "Contrastive Search" [8]. Hyperparameters used for STEER: $\gamma = 0.2, \eta = 0.4$, no. negative prompts $= 10$.

Win rate against other sampling methods

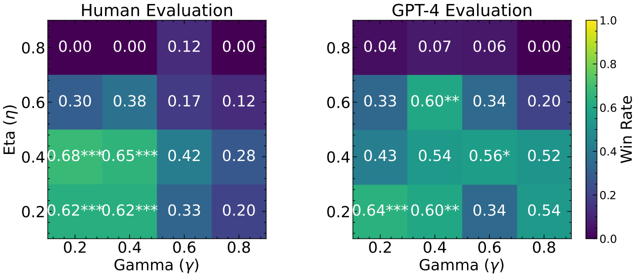


Figure: Win rate of STEER against nucleus sampling in the hypothesis generation task. The levels of significance are marked as follows: $***$ denotes $p < 0.001$, $**$ denotes $0.001 \leq p < 0.01$, and $*$ denotes $0.01 < p \leq 0.05$.

Sensitivity analysis

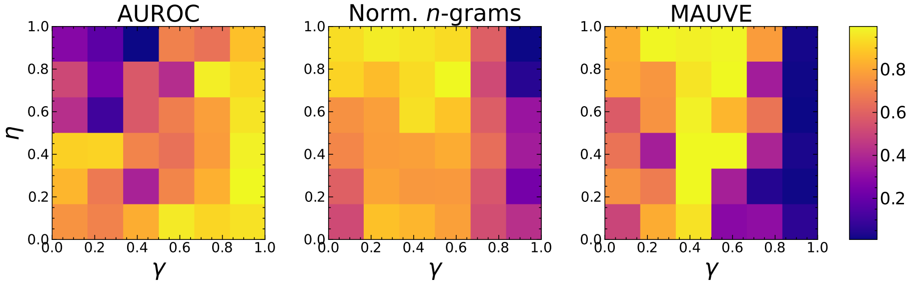


Figure: Performance of Falcon-7B on the hypothesis generation task when varying the contrastive guidance hyperparameter γ and the negative prompting hyperparameter η. 50 examples were produced for each combination of γ and η to evaluate the metrics on. A lower AUROC is better, and higher normalised n-grams and MAUVE are better.

Ablations (of a kind)

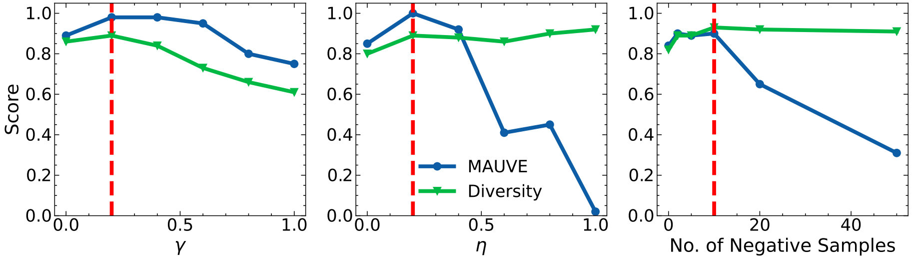


Figure: Trade-offs when varying one hyperparameter at a time, keeping the other fixed at 0 (for γ and η, which is not necessarily the optimal value). For the number of negative prompts, we set $(\gamma, \eta) = (0.4, 0.4)$. The dashed red vertical line shows the point at which the sum of MAUVE and diversity score is greatest.

Downstream accuracy

- For Jigsaw toxic comments and CommonsenseQA, we can generate synthetic data and train a model on a downstream task e.g. *text classification*. This is a test of knowledge distillation
- For the Arxiv Hypotheses, we can examine the win-rate of different generation methods against the real data using expert evaluators

	STEER	Greedy	Nucleus	Contrastive	*Real*
Jigsaw	**0.94** ± 0.02	0.91 ± 0.03	0.90 ± 0.02	0.89 ± 0.01	*0.98*
QA	**0.41** ± 0.03	0.35 ± 0.04	0.40 ± 0.03	0.29 ± 0.02	*0.55*

Figure: Downstream accuracy comparison for Falcon-7B across two datasets: Jigsaw Toxic Comments and CommonsenseQA. Models were evaluated on five different splits of the real data.

Illustrating the trade-off

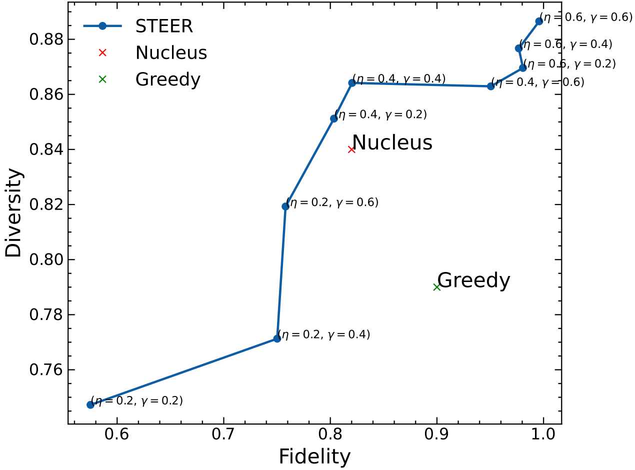


Figure: Trade-off between MAUVE score and normalised *n*-grams score for 50 STEER generations in each hyperparameter combination.

UMAP and convex hull precision/recall

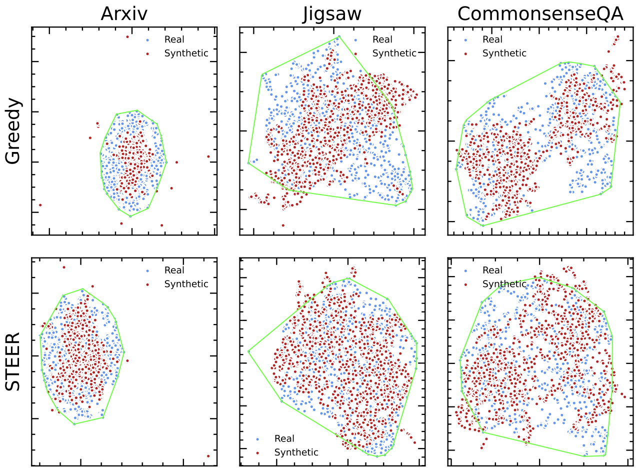


Figure: UMAP visualisations of the embeddings for real and synthetic data, with the real embeddings colored in blue and the synthetic ones in red. The convex hull surrounding the real data is delineated by the green line.

UMAP and convex hull precision/recall

$$\text{Convex Hull Precision} = \frac{|\{X_{s,j} \mid X_{s,j} \in \mathcal{H}_r\}_{j=1}^{m}|}{m} \qquad \text{Convex Hull Recall} = \frac{|\{X_{r,i} \mid X_{r,i} \in \mathcal{H}_s\}_{i=1}^{n}|}{n}$$

	Convex Hull Precision	Convex Hull Recall	F-score
ArXiv Hypotheses			
Greedy	**0.997**	0.949	0.972
Nucleus	0.996	0.952	0.974
Contrastive	0.996	0.867	0.927
STEER	0.994	**0.963**	**0.978**
Jigsaw Toxic			
Greedy	0.785	0.910	0.843
Nucleus	**0.802**	0.807	0.805
Contrastive	0.733	0.919	0.815
STEER	0.772	**0.993**	**0.869**
CommonsenseQA			
Greedy	0.886	0.969	0.926
Nucleus	**0.945**	0.953	**0.949**
Contrastive	0.930	0.9610	0.945
STEER	0.878	**0.979**	0.926

Experts don't like it as much...

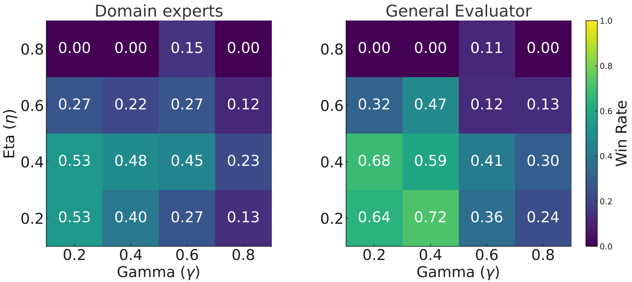


Figure: Comparing the win rates of STEER vs nucleus for astronomy domain experts, defined by having postdoctoral qualifications in astronomy (three evaluators) compared with general annotators (five evaluators).

Table of Contents

Why do we care?

- The “GPU poor” can’t afford to fine-tune Falcon-180B; use a smaller model and boost it with STEER
- Generate diverse synthetic datasets for recursively improving language models [9]
- Quantitative way to gain control without subjective/qualitative prompt engineering
- This work itself is possibly outdated (GPT-4 hits the Pareto curve). But there is a lot of potential for work which lets us choose which part of the curve we want to be on

Moving from logit to latent space

- Instead of subtracting the **logits**, subtract the **weights** of FT model from base model
- This gives a *task vector*, such that moving in the direction of the vector improves performance on the task the FT model is good at
- Can also utilise negation, addition, and even transitive properties to linearly "steer" the model in the weight space

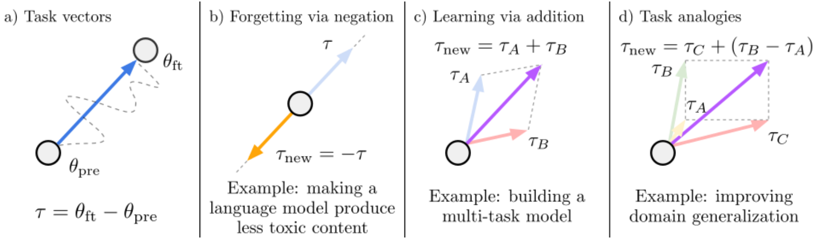


Figure: The figure and idea come from the *Editing models with task arithmetic* paper by Ilharco et. al [3].

Wrapping up

- Hopefully the logic behind this method provides some inspiration for other LLM-based challenges; can we *generalise* this method of attracting and repelling?
- What other types of subtraction can you come up with that might improve performance? (Subtracting unconditional distributions, logits of a terrible model, logits from short vs. long context, etc.)
- How do we evaluate these things? Our synthetic metrics can be "hacked", as seen from Figure 13
- Thanks to my supervisor Thang Bui and the wonderful people from universeTBD, particularly Yuan-Sen Ting and Jo Ciuca

References I

[1] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. arXiv:2105.05233 [cs, stat]. June 2021. DOI: 10.48550/arXiv.2105.05233. URL: http://arxiv.org/abs/2105.05233 (visited on 07/10/2023).

[2] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. arXiv:2207.12598 [cs]. July 2022. URL: http://arxiv.org/abs/2207.12598 (visited on 07/10/2023).

[3] Gabriel Ilharco et al. "Editing models with task arithmetic". In: *arXiv preprint arXiv:2212.04089* (2022).

[4] Xiang Lisa Li et al. *Contrastive Decoding: Open-ended Text Generation as Optimization*. arXiv:2210.15097 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.15097. URL: http://arxiv.org/abs/2210.15097 (visited on 07/07/2023).

[5] Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. *Coherence boosting: When your pretrained language model is not paying enough attention*. 2022. arXiv: 2110.08294 [cs.CL].

[6] Guillaume Sanchez et al. *Stay on topic with Classifier-Free Guidance*. arXiv:2306.17806 [cs]. June 2023. DOI: 10.48550/arXiv.2306.17806. URL: http://arxiv.org/abs/2306.17806 (visited on 07/06/2023).

[7] Weijia Shi et al. *Trusting Your Evidence: Hallucinate Less with Context-aware Decoding*. arXiv:2305.14739 [cs]. May 2023. DOI: 10.48550/arXiv.2305.14739. URL: http://arxiv.org/abs/2305.14739 (visited on 07/07/2023).

[8] Yixuan Su et al. *A Contrastive Framework for Neural Text Generation*. 2022. arXiv: 2202.06417 [cs.CL].

[9] Jerry Wei et al. "Simple synthetic data reduces sycophancy in large language models". In: *arXiv preprint arXiv:2308.03958* (2023).