

Coherence boosting: When your pretrained language model is not paying enough attention

Nikolay Malkin
Mila / Université de Montréal
nikolay.malkin@mila.quebec

Zhen Wang
Ohio State University
wang.9215@osu.edu

Nebojsa Jojic
Microsoft Research
jojic@microsoft.com

Abstract

Long-range semantic coherence remains a challenge in automatic language generation and understanding. We demonstrate that large language models have insufficiently learned the effect of distant words on next-token prediction. We present *coherence boosting*, an inference procedure that increases a LM’s focus on a long context. We show the benefits of coherence boosting with pretrained models by distributional analyses of generated ordinary text and dialog responses. It is also found that coherence boosting with state-of-the-art models for various zero-shot NLP tasks yields performance gains with no additional training.

1 Introduction

Language models (LMs) are commonly evaluated for their ability to generate, rank, or classify coherent spans of text. **Long-range semantic coherence** is a unifying feature of modern NLP benchmarks and applications, whether they are about producing short answers to questions, ranking answer choices by their consistency with world knowledge, or generating long responses.

Large nonspecialized LMs, such as GPT-2 and -3 (Radford et al., 2019; Brown et al., 2020), sometimes fail to understand or use the semantic link between a text and its prompt or long-range context (Fig. 1). Samples from these LMs have an unnaturally low density of words that require many tokens of context to predict (§4.1), and the scores that the models give to completions of prompts indicate that they are oversensitive to recent context (§5).

We hypothesize that these failures arise from modeling choices and distribution shift. Specifically, autoregressive LMs are typically fit to a multi-objective problem: simultaneously maximizing token likelihoods conditioned on many lengths of truncated context (§2.1). Yet, at generation or

scoring time, likelihoods are conditioned on the entire prompt or previously generated string, specifically selected to be coherent or even guaranteed to influence the output. The two common solutions – finetuning models on one or multiple tasks (Khashabi et al., 2020; Sanh et al., 2022) and improving models or prompts to facilitate in-context learning (Brown et al., 2020; Schick and Schütze, 2021) – do not directly target the problem of long-range coherence.

This paper proposes **coherence boosting**, a simple inference-time procedure that increases the effect of distant words on predicted token distributions and is applicable in both generation and ranking settings. A pretrained model is viewed as an *ensemble* of experts that produce token distributions conditioned on varying lengths of context. These experts are log-linearly mixed to form a predictor that is superior to the base model (§2).

Coherence boosting greatly improves prediction of words that depend on a long context, as evidenced by state-of-the-art results on tasks specially meant to assess models’ attention to distant words (§3). In generation of generic text and dialog responses, we show that coherence boosting brings the frequency of occurrence of such words close to that seen in natural text (§4). Beyond generation, we study diverse multiple-choice tasks (§5), in which examples are known to be highly coherent. Coherence boosting does not modify the base model and depends on a single parameter than can be estimated in one pass through a validation set, yet is a competitive adaptation algorithm.

1.1 Background and related work

Balance between satisfaction of short-range statistical constraints and maintenance of long-range structure was a central question of language generation long before neural language modeling. To compensate for the sparsity of the learning signal for long-range influences, *n*-gram models and

Code: github.com/zhenwang9102/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 w

$p_{\text{full}} = f(w \mid \text{full})$ 1. Alex (1.9%) 2. **Natasha** (1.7%) 3. also (1.5%)
 $p_{\text{short}} = f(w \mid \text{short})$ 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 w

$p_{\text{full}} = f(w \mid \text{full})$ 1. sense (9.0%) 2. in (2.0%) 3. . (1.9%) ... 13. **metre** (0.6%)
 $p_{\text{short}} = f(w \mid \text{short})$ 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 w

$p_{\text{full}} = f(w \mid \text{full})$ 1. bar (4.8%) 2. drink (3.7%) 3. with (3.5%) ... 13. **stout** (2.7%)
 $p_{\text{short}} = f(w \mid \text{short})$ 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 w

$p_{\text{full}} = f(w \mid \text{full})$ 1. important (5.6%) 2. bad (4.6%) 3. **debilitating** (4.3%)
 $p_{\text{short}} = f(w \mid \text{short})$ 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 1: 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 of the distributions makes the correct word most likely. Sampling from such a mixture at each generation step (*coherence boosting*) improves the quality of output text (§4). (Dialog example constructed by the authors; other examples from OpenWebText.)

early neural language models used ‘backing-off’ schemes that interpolate between predictors with different context lengths (Chen and Goodman, 1996; Bengio et al., 2003). Neural language modeling brought a need for recurrent units with better numerical properties for propagating information over long distances (Hochreiter and Schmidhuber, 1997; Cho et al., 2014) and eventually saw the reintroduction of alignment variables (Brown et al., 1993) into generation in the form of attention (Bahdanau et al., 2015; Vaswani et al., 2017). Attention is at the core of Transformer LMs, including GPT.

Language models are being trained on and adapted to ever-longer input sequences (Beltagy et al., 2020; Zaheer et al., 2020; Roy et al., 2021; Press et al., 2022), but they remain undersensitive to distant content or syntax (Khandelwal et al., 2018; Sun et al., 2021) and are easily fooled by recency bias in few-shot prompts (Zhao et al., 2021) or multi-turn conversations (Sankar et al., 2019).

Recent work has continued to study inference-time procedures that prevent text sampled from LMs from degenerating into nonsense. Most of these procedures, such as tempered sampling and top- k /top- p truncation (Fan et al., 2018; Holtzman et al., 2019), independently modify the output dis-

tribution at each generation step to decrease its entropy and diminish its low-likelihood tail. Holtzman et al. (2019) and Meister and Cotterell (2021) found that such local modifications increase the quality of long generated sequences; we adopt and extend their methodology in §4.1.

For dialog systems, Li et al. (2016) propose a decoding scheme that maximizes a mutual information criterion, which explicitly optimizes for dependence of generated text on prompts – a special case of coherence boosting. In multiple-choice tasks, where a model must choose one of several given completions of a prompt, Brown et al. (2020) observe that selecting the completion that maximizes the conditional likelihood of the completion following the prompt often favors completions having high *unconditional* likelihood (likelihood following an empty or dummy prompt) and, for some tasks, chooses to divide the scores of candidate answers by their unconditional likelihoods. This is also a special case of coherence boosting.

Such scoring modifications are more thoroughly studied by Zhao et al. (2021); Holtzman et al. (2021). The latter attributes the problem to ‘surface form competition’: there are many variants of the correct completion that together may capture a

large part of probability mass, but the form of the given answer choice alone is not the most likely. However, we show that other causes are at play: surface form competition is impossible when the completion is known to be a single token and the range of choices is the whole vocabulary (§3), and it is not applicable to open-ended generation (§4).

2 Coherence boosting

In this section, f is an autoregressive LM over a vocabulary V with learnable parameters θ , taking as input a variable number of tokens (up to a maximum context length M) and producing a vector of next-token likelihoods:

$$f(w_1, \dots, w_n; \theta) \in \Delta(V), \quad w_1, \dots, w_n \in V,$$

where $\Delta(V)$ is the probability simplex over V . We will write the w -th component of this output vector as a conditional likelihood, $f(w | w_1, \dots, w_n; \theta)$.

We denote by f_k the model evaluated on only the *last* k input tokens, ignoring earlier tokens:

$$f_k(w_1, \dots, w_n; \theta) := f(w_{n-k+1}, \dots, w_n; \theta).$$

Coherence boosting for next-token prediction.

Coherence boosting for a model f selects real-valued weights $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_M)$ and produces a new language model f_α , defined by

$$f_\alpha(w_1, \dots, w_n; \theta) := \text{softmax} \left(\sum_{k=1}^M \alpha_k \log f_k(w_1, \dots, w_n; \theta) \right), \quad (1)$$

where \log is taken element-wise, or, equivalently,

$$f_\alpha(w | w_1, \dots, w_n; \theta) \propto \prod_{k=1}^M f_k(w | w_1, \dots, w_n; \theta)^{\alpha_k}.$$

This is a weighted product-of-experts model, where the ‘experts’ are copies of the base model f evaluated on different context lengths.

Because evaluating f is expensive, we use sparse weights α , as the expression (1) depends only on those f_k for which $\alpha_k \neq 0$. In Fig. 1 and in the experiments, we allow α to have only two nonzero entries: when computing likelihoods of words following a sequence of length n , we consider weighted products of $f_{\max} := f_n$ (the full context) and an f_k with $k \leq n$ (a short context, either of fixed length or decided by prompt structure as in §4.2).

As its name suggests, the form of coherence boosting in (1) bears a resemblance to log-linear

boosting for multiclass classification (Friedman et al., 2000). However, our weak classifiers are pretrained and share all of their parameters, not obtained by an iterative procedure of training on reweighted data, and we permit negative weights.¹

Coherence boosting for answer selection. In multiple-choice problems, a LM must choose the best answer following a context, which consists of a premise or passage followed by a shorter *premise-free context* (either a short phrase, such as “Answer:”, that incites the LM to generate an answer in the right format, or a hypothesis that depends on the premise). The full context is the concatenation of the premise and the premise-free context (§E).

By the autoregressive factorization, the model f assigns conditional likelihoods to *sequences* of tokens following context. A typical model for answer selection ranks the candidate answers a_i (sequences of tokens) by $f(a_i | \text{full context}; \theta)$ and outputs the highest-ranked a_i . *Coherence boosting* chooses a parameter α and ranks the choices by:

$$\begin{aligned} & \log f(a_i | \text{full context}; \theta) + \\ & + \alpha \log f(a_i | \text{premise-free context}; \theta). \end{aligned} \quad (2)$$

This is a log-linear combination of two models: f evaluated with full context and with a partial context. When $\alpha = 0$, ranking by (2) is equivalent to ranking by the base model. When $\alpha = -1$, it is equivalent to dividing the base model’s score by the score of each answer conditioned on the prompt (short context), and thus to maximizing pointwise mutual information between the premise and the answer conditional on the premise-free context. Unlike Brown et al. (2020); Holtzman et al. (2021), our formulation allows the premise-free context to include information specific to the example, not only a domain-specific dummy prompt.

We expect coherence boosting to correct for an oversensitivity to the premise-free context, and thus the optimal α will typically be negative (see §5).

2.1 Why should boosting models be better than full-length predictors?

Multi-objective training. As we will now see, the training of the model f simultaneously fits all of

¹As for the first half of the term ‘coherence boosting’, Howcroft et al. (2020); Belz et al. (2020) found that very incoherent definitions of the word ‘coherence’ abound in the natural language evaluation literature. The reader is asked to forgive us for the loose definition of ‘long-range semantic coherence’ adopted in this paper.

the predictors f_k , which share parameters θ . Each training iteration samples a sequence (or batch of sequences) of a chosen maximum length $M + 1$ from the data distribution \mathcal{D} and minimizes the average negative log-likelihood (NLL) of *all* words following the parts of the sequence that precede them: the optimization criterion is:

$$\mathbb{E}_{w_1 \dots w_{M+1} \sim \mathcal{D}} \frac{1}{M} \sum_{k=1}^M -\log f(w_{k+1} | w_1, \dots, w_k; \theta).$$

If \mathcal{D} is uniform over all length- $(M + 1)$ subsequences of a training corpus, any given word is equally likely to appear in all positions within a sampled sequence², and the criterion is equal to

$$\sum_{k=1}^M \frac{1}{M} \underbrace{\mathbb{E}[-\log f_k(w_{M+1} | w_1, \dots, w_M; \theta)]}_{\mathcal{L}_k(\theta)}, \quad (3)$$

This is a uniform scalarization of an M -task problem: the k -th objective $\mathcal{L}_k(\theta)$ is the expected NLL of a word in the corpus following k context words.

This situation is different from that seen at *generation* time. If the text generated so far is $w_1 w_2 \dots w_n$, the distribution from which the next word w_{n+1} is sampled is $f_n(w_1, \dots, w_n; \theta)$ – only the ensemble member using full context is used. However, if the string $w_1 \dots w_n w_{n+1}$ had been seen in training, f would have been trained to predict w_{n+1} given *all partial contexts*, with equal weight given to all prediction losses. Thus, f is trained to make predictions on data it never sees in evaluation, and may be prevented from optimally learning to use long context: parameters that locally optimize (3) are locally Pareto-optimal for the *set* of prediction losses $\mathcal{L}_1, \dots, \mathcal{L}_M$, but not necessarily optimal for any individual \mathcal{L}_k . An ensemble of the f_k ($k \leq n$) may be a better predictor than f_n alone. (See §A for further analysis of when this occurs.)

Undertraining. The parameters θ are shared by the predictors f_k , and modeling power must be spread among the losses $\mathcal{L}_k(\theta)$. The short-context predictors are easier to fit, while sequences in which long context affects the prediction are rare. We expect sensitivity to long context, and precision in modeling its effect, to be especially diminished if the model is undertrained.

²Many authors leave unspecified the way in which training batches are formed from a corpus of input documents. Here we assume that all training documents are concatenated into one (very long) document separated by end-of-text tokens and ignore minute effects near the start and end of this document.

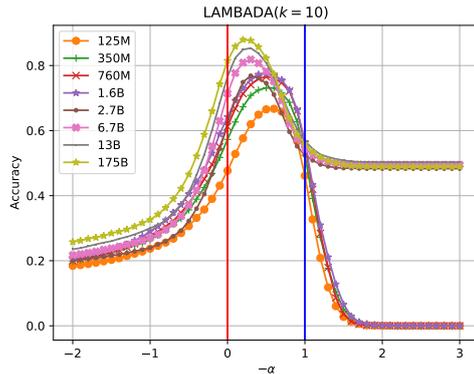


Figure 2: Model comparison on LAMBADA with $k = 10$ and varying α_k . The red line ($\alpha = 0$) is the base LM f_{\max} . (The different right tails of GPT-3 models are due to top-100 truncation of logits returned by the API.)

Distribution shift. While the training procedure causes a bias against the influence of longer contexts on generation, we see the opposite bias in downstream tasks (question answering, natural language inference, adversarial probes for common sense): Many modern NLP benchmarks try to challenge models to use long context (§3, §5).

3 Experiments: LAMBADA

The LAMBADA dataset (Paperno et al., 2016) tests LMs’ understanding of long-range dependencies by measuring the prediction of the final words in passages of several sentences. The task explicitly requires reasoning over a broad context: humans can reliably guess the last word when given a whole passage, but not when given only the last sentence.

We perform experiments with the GPT family of models, closely replicating the evaluation setting of Radford et al. (2019).³ We predict the final word as the top-ranked token under the boosted model $f_{\max} f_k^{\alpha_k}$, where f_{\max} is the model taking the full available context and k, α_k are the chosen length and coefficient of the short context. To choose k and α_k , we do a grid search on the validation set and apply the best values to the testing set.

Results. Table 1 shows the accuracies and optimal parameter values k^*, α_k^* . Coherence boosting vastly reduces prediction error for all models. In particular, the boosted GPT-2 Small performs better than the original GPT-3 2.7B. The boosted GPT-3 175B achieves a new state of the art.

³Certain details are omitted by Radford et al. (2019). Based on <https://github.com/openai/gpt-2/issues/131>, we nearly match baseline accuracy by predicting the last subword token, rather than the last word.

	GPT-2				GPT-3			
	125M	350M	760M	1.6B	2.7B	6.7B	13B	175B
f_{\max}	47.66	57.29	61.23	64.25	62.39	71.40	76.58	81.51
CB ($\alpha_k = \alpha_k^*$)	66.70	73.53	76.54	77.53	77.00	81.84	86.36	88.61
α_k^*	-0.6	-0.5	-0.5	-0.5	-0.3	-0.3	-0.3	-0.2
k^*	10	11	10	9	9	10	3	3

Table 1: Accuracy (%) and optimal boosting parameters on LAMBADA: f_{\max} is the full-context model without boosting; CB is our model with the optimal boosting parameters (last two rows).

Other than the impressive performance gain, we highlight two observations. **(1)** The optimal α_k is always negative, indicating that the optimal mixture of models penalizes the influence of short-range context relative to long-range context. **(2)** With increasing model size, the optimal α_k and k become closer to 0. This means that bigger models capture long-range coherence better than small models, as they have less need to penalize the effect of short context. (Fig. 2 shows the accuracy curves for all models by sweeping α_k with a fixed k . The peak clearly moves to the left as model size grows.)

4 Experiments: Language generation

4.1 Generic text

The experiment in this section extends that of Holtzman et al. (2019). A selection of 5000 articles from WebText (Radford et al., 2019) is taken as a reference corpus of human-written text. A language model (for us, GPT-2 Large) is prompted to generate text conditioned only on the *first sentence* of each of these articles, up to a maximum of 200 tokens, yielding 5000 machine-generated texts.

The human-written and machine-generated texts are compared by four automatic metrics: **perplexity** under the base LM, **self-BLEU-4** (Zhu et al. (2018); the mean BLEU-4 score of a generated text with respect to all other generated texts as references), **Zipf coefficient** (the linear regression coefficient between log-rank and log-frequency of generated tokens) and **repetition** (the fraction of generated texts that end in a repeating sequence of tokens). It is desirable for a model and inference procedure to produce text that is as close as possible in these metrics to the human-written reference.

To measure long-range semantic coherence in the generated text, we define three new metrics:

Long-range repetition (LR_n): For a whole number n and document D , let $S(D)$ be the number of distinct tokens in D , and let $R_n(D)$ be the number

of distinct tokens for which the distance between their first and last occurrence in D is at least n positions. The long-range repetition score LR_n of a corpus $\{D_1, \dots, D_{5000}\}$ is a macro-average:

$$LR_n := \frac{\sum_{i=1}^{5000} R_n(D_i)}{\sum_{i=1}^{5000} S(D_i)}.$$

This simple measure of lexical coherence favors repetition of words long after they are first used, but gives lower weight to documents that degenerate into repetition of a short span.

Long-dependent token frequency (LTF): A *long-dependent token* is one to which the base LM assigns a likelihood of at least 20% given its full context, but a likelihood of less than 5% given only the 20 tokens of context preceding it. We compute the frequency of long-dependent tokens among all generated tokens.

Long-short likelihood difference (δ): The mean difference in likelihoods assigned to tokens by the base LM conditioned on full context and conditioned on 20 tokens of context.

Although some choices of constants are needed to define LTF and δ , we intend them to be intuitive summaries of long-range coherence in the absence of established metrics. In particular, 20 tokens is close to the length of one sentence in typical English text.

We sample 5000 document completions from GPT-2 Large following sampling procedures with a range of boosting schemes. We consider models of the form $f_k^{\alpha_k} f_{\max}^{1-\alpha_k}$, for $k \in \{8, 16, 32, 64\}$ and $\alpha_k \in \{-0.4, -0.2, -0.1, -0.05, -0.025, 0\}$. (Such a parametrization of boosting parameters was chosen to ensure that when the context has length less than k – or the distant context has very little effect on the next word – the boosted model becomes equivalent to the untempered f_{\max} .) Top- p truncation with $p = 0.95$ is applied to all models.

Inference method	from Holtzman et al. (2019)				lex coherence		long-dep tokens	
	ppl	BLEU-4	Zipf	rep %	LR ₅₀ %	LR ₁₀₀ %	δ %	LTF %
Sampling	23.53	0.28	0.93	0.22	12.92	7.71	4.87	3.28
Sampling ($T = 0.9$)	10.60	0.35	0.96	0.66	16.36	10.01	6.54	4.15
Nucleus ($p = 0.95$)	13.48	0.32	0.95	0.46	15.06	9.11	5.65	3.62
+ boost ($k = 32, \alpha_k = -0.05$)	12.81	0.31	<i>0.94</i>	0.34	<i>15.54</i>	9.42	<i>6.16</i>	3.98
+ boost ($k = 64, \alpha_k = -0.1$)	12.93	<i>0.32</i>	<i>0.95</i>	<i>0.46</i>	15.75	<i>9.67</i>	<i>6.10</i>	<i>3.95</i>
+ self-tune (§B)	10.16	0.33	0.95	0.64	16.19	9.85	6.59	4.16
Human	13.19	0.31	0.93	0.28	15.95	9.51	6.54	4.03

Table 2: Distributional metrics of WebText completions. The last four columns are measures of long-range coherence (§4.1). (Nearest-to-human values in **bold**, boosting models better than top- p sampling alone in *italics*.)

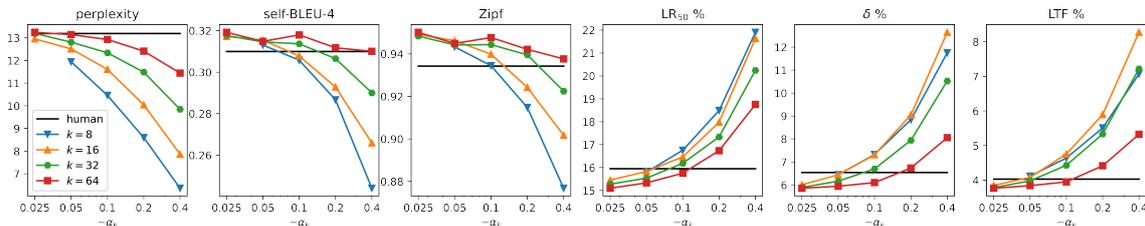


Figure 3: Effect of k and α_k on metrics from Table 2. The horizontal line marks the score of the human reference.

Results. Metrics of two of the best models, with $k = 32, \alpha_k = -0.05$ and $k = 64, \alpha_k = -0.1$, are shown in Table 2. In particular, the latter model generates text that is closer to the human reference, or equally close, to the pure top- p sampling ($\alpha_k = 0$) baseline in all metrics, with the greatest improvement seen in the coherence measures.

Fig. 3 shows the dependence of selected metrics on k and α_k . Coherence boosting brings all metrics closer to those of human text. As k increases, the optimal α_k grows in magnitude. This is expected: the predictive effect of tokens more than k positions away decreases with k (f_k approaches f_{\max}).

We also note that a simple sampling with temperature 0.9 performs better than top- p sampling in most of the coherence metrics. This suggests that the improvements accomplished by top- p truncation come at the cost of introducing a bias towards tokens that are predictable from a short context. Coherence boosting corrects this bias without sacrificing the gains in other measures.

An example of human, top- p , and coherence boosting outputs is shown in Table D.1.

4.2 Dialog systems

This experiment is based on the Dialog System Technology Challenge 7 (DSTC7) (Galley et al., 2019), which benchmarks generation of dialog re-

sponses conditioned on one or more turns of conversation context. As a base model, we use DialogPT (Zhang et al., 2020c), a GPT-2 Small variant that demonstrated strong results on this task.

Dialog systems’ responses to the 2208 conversation prompts⁴ are scored against human-written reference responses (five for each example). Following Zhang et al. (2020c), we use the n -gram overlap metrics **NIST** (Doddington, 2002), **BLEU** (Papineni et al., 2002), and **METEOR** (Lavie and Agarwal, 2007), as well as two intrinsic measures of n -gram diversity from Li et al. (2016); Zhang et al. (2018): **Distinct- n** and **Entropy- n** . It is desirable for a dialog system to reach scores close to those of the human responses in all metrics.

In addition to the decoding algorithms considered by (Zhang et al., 2020c) – beam search and greedy decoding – we consider greedy decoding with a coherence boosting model. As long and short predictors, we use DialogPT conditioned on the full conversation context and on *only the (context-free) response generated so far*. That is, if the conversation context is S and the text generated so far is $w_1 \dots w_k$, then w_{k+1} is predicted using the model $f_{\max} f_{k+1}^\alpha$, evaluated on the string $S \langle \text{sep} \rangle w_1 \dots w_k$, where $\langle \text{sep} \rangle$ is the turn separa-

⁴The DSTC7 evaluation data, scraped from Reddit, is undisclosed; we reacquire it using officially released code.

Inference method	NIST		BLEU		METEOR	diversity metrics			avg len
	N-2	N-4	B-2	B-4		Ent-4	Dist-1	Dist-2	
Beam ($b = 10$)	0.02	0.02	12.81	3.23	5.35	6.06	14.03	34.59	5.81
Greedy	1.62	1.63	9.92	1.72	6.78	6.45	6.19	17.56	13.30
+ boost ($\alpha = -0.3$)	0.72	0.73	13.82	3.53	6.91	8.54	16.81	49.35	9.75
+ boost ($\alpha = -0.7$)	1.78	1.79	6.33	0.94	5.55	9.78	28.00	72.46	16.63
Human	2.63	2.65	12.36	3.13	8.31	10.44	16.65	67.01	18.73

Table 3: Metrics of DialoGPT responses on DSTC7. Nearest-to-human values in each column are **bolded**.

tor token. We consider $\alpha \in \{0, -0.1, \dots, -0.8\}$.

Results. Table 3 shows the metrics of the boosting models that reach the peak average NIST and BLEU scores ($\alpha = -0.3$ and $\alpha = -0.7$). Increasing the magnitude of α leads to responses that are more relevant to the prompt (higher BLEU and NIST) and more diverse than those from greedy decoding. As $-\alpha$ grows large, the boosting model favors creative responses that are relevant to the prompt (high NIST), but simple responses that are common in the reference data become unlikely (low BLEU).⁵

We observed that the responses with $\alpha = -0.7$, despite the superior metrics, are more likely to be ungrammatical and innovate words in an effort to use tokens relevant to the prompt. In practice, improving dialog systems with coherence boosting may require techniques to prevent these side effects, such as repetition penalties or relaxation of greedy decoding to low-temperature sampling.

Finally, we note that the learning of DialoGPT was initialized with a pretrained GPT-2 and uses GPT-2’s end-of-text token as the turn separator. This choice may reduce DialoGPT’s attention to past turns, as tokens *preceding* the end-of-text token are never informative in GPT-2’s training data.

5 Experiments: Language understanding

We evaluate coherence boosting on zero-shot language understanding and inference tasks, where examples are expected to be highly coherent.

We study 15 datasets in 5 categories of tasks. **(1) Cloze tasks:** *StoryCloze* (Mostafazadeh et al., 2016), *HellaSwag* (Zellers et al., 2019), and *COPA* (Roemmele et al., 2011). **(2) Question answering:** *CommonsenseQA* (CsQA) (Talmor et al., 2019), *OpenBookQA* (OBQA) (Mihaylov et al.,

2018), *ARC Easy / Challenge* (ARC-E/C) (Clark et al., 2018), and *PIQA* (Bisk et al., 2020). **(3) Text classification:** *SST-2/5* (Socher et al., 2013), *TREC* (Voorhees and Tice, 2000), *AGNews* (Zhang et al., 2015). **(4) Natural language inference:** *RTE* (Dagan et al., 2005), *CB* (De Marneffe et al., 2019), and *BoolQ* (Clark et al., 2019). **(5) Fact knowledge retrieval:** *LAMA* (Petroni et al., 2019).

All tasks except LAMA are formulated as multiple-choice problems. We convert text classification and inference tasks to multiple-choice tasks by choosing meaningful answer words, e.g., “True”/“False”. The prediction is made by selecting the choice with the highest LM likelihood.

For in-context learning of GPT models, prompt formats greatly impact performance. We follow previous work (Brown et al., 2020; Zhao et al., 2021; Holtzman et al., 2021) to create natural prompts to enlarge the effectiveness of in-context learning, but we do not aim to optimize the full and context-free prompt format: our goal is to evaluate coherence boosting models with a fixed prompt. The prompt formats we use are listed in Table E.1. As described in §2, within each prompt we identify a *premise-free context*, which is used as the context for the short-range model in coherence boosting.

For each dataset, we pick the optimal value α^* of the parameter α on the validation set and report the accuracy on testing set. (If no testing set is publicly available, we choose α on a subset of the training set and report the final number on the validation set.) Across all experiments, we do not put any few-shot examples in the prompt.

For the knowledge retrieval task, we follow Zhao et al. (2021)’s data split of LAMA and evaluate GPT models on facts whose missing answers are at the end of the sentence (to fit the nature of autoregressive language models). We limit the prompt length to be larger than 5 tokens and rerun the model from Zhao et al. (2021) on the new data.

⁵Galley et al. (2019) argue that NIST and diversity metrics are more informative measures than BLEU for multi-reference scoring, since BLEU favors systems that often produce responses with little relation to the prompt (e.g., “I don’t know”).

	GPT-2 Small (125M)				GPT-2 XL (1.6B)				GPT-3 175B			
	f_{\max}	$\alpha = -1$	$\alpha = \alpha^*$	α^*	f_{\max}	$\alpha = -1$	$\alpha = \alpha^*$	α^*	f_{\max}	$\alpha = -1$	$\alpha = \alpha^*$	α^*
StoryCloze	59.91	64.78	64.24	-1.02	67.56	75.09	76.75	-0.69	79.16	82.90	86.85	-0.64
HellaSwag	28.92	30.99	31.84	-0.90	40.00	42.60	47.66	-0.78	59.18	62.66	72.35	-0.76
COPA	62.00	56.00	64.00	-0.69	73.00	70.00	77.00	-0.44	93.00	87.00	94.00	-0.52
CsQA	29.48	42.26	43.16	-0.81	37.84	50.45	52.91	-0.75	61.10	67.98	70.43	-0.68
OBQA	11.20	30.60	40.80	-1.62	15.60	38.40	47.00	-1.88	28.00	52.20	52.60	-1.09
ARC-E	43.81	42.09	46.00	-0.34	58.29	51.43	60.31	-0.36	76.22	69.19	78.32	-0.44
ARC-C	19.03	26.11	29.10	-4.19	25.00	33.53	34.39	-1.14	43.94	50.60	49.23	-1.08
PIQA	62.89	57.45	63.44	-0.61	70.84	60.45	71.49	-0.43	79.27	66.32	78.94	-0.60
SST2	65.68	74.74	82.32	-2.22	86.38	84.51	86.93	-0.09	86.16	88.14	89.84	-0.54
SST5	25.93	30.90	30.90	-1.20	28.69	38.73	36.92	-1.69	31.22	34.75	38.51	-1.39
AGNews	58.55	60.78	62.20	-0.62	67.17	67.43	68.26	-0.40	71.66	71.74	71.75	0.16
TREC	23.40	29.60	32.20	-0.80	23.40	27.40	40.00	-0.79	52.40	47.00	56.00	-0.56
BoolQ	49.36	58.07	62.14	-3.04	62.14	63.46	63.21	-0.64	71.56	73.70	72.69	-0.39
RTE	51.26	49.82	53.79	-0.30	49.10	48.74	49.10	0.90	55.96	57.40	60.29	-0.60
CB	12.50	23.21	48.21	-2.40	30.36	51.79	66.07	-1.90	5.36	25.00	28.57	-1.91
Average	40.26	45.16	50.29	-1.39	49.02	53.60	58.53	-0.74	59.61	62.44	66.69	-0.73

Table 4: Testing accuracy (%) of three representative GPT models on multiple-choice tasks. The first column for each model is the full-context model, the second is our model only when $\alpha = -1$ (a baseline), and the third column is our model with the optimal α chosen on a validation set. The fourth column shows this optimal value of α .

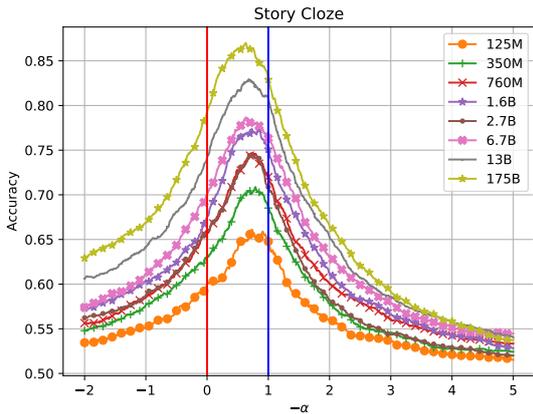


Figure 4: Model comparison for the StoryCloze task. The red line $\alpha = 0$ indicates the base model, and the blue line $\alpha = -1$ is an unconditional normalization. See Figs. F.1 and F.2 for plots for other tasks, and note that they do not all have the same shape.

Results: Multiple-choice tasks. Results of three representative base models on all multiple-choice tasks are presented in Table 4. (Results for all models are in Tables F.1 and F.2.) We compare our best model with two baselines, $\alpha = 0$ (f_{\max}) and $\alpha = -1$. The former one is the original full-context model, while the latter is, for most tasks, a form of unconditional probability normalization as performed by Brown et al. (2020); Holtzman et al. (2021). We also compare our best model with other inference methods (Holtzman et al., 2021;

Min et al., 2021) in Tables F.3 and F.4.

By comparing the third column with the first two columns within each model in Table 4, we can see that our method with the selected α generally improves the accuracy on all tasks. Some of the improvements are dramatic, where boosted GPT-2 Small outperforms GPT-2 XL’s base model (e.g., CsQA, OBQA, ARC-C) and is even comparable with GPT-3 175B’s base model (e.g., SST-2, SST-5, RTE). We make similar conclusions when comparing coherence boosting with other inference methods in Tables F.3 and F.4.

We observe that the optimal α depends on tasks and models (fourth column within each model), which means that α cannot be heuristically set to 0 or -1 as in past work. This finding suggests the necessity of searching for an optimal α . We visualize the accuracy curve by varying α in the testing set of all datasets. We show the curve for StoryCloze in Fig. 4 and present similar figures for all tasks in Figs. F.1 and F.2.

Consistent with the results on LAMBADA (§3), the optimal α is usually negative, and its absolute value tends to decrease with the model size. We selected the optimal α by the validation set, but future work may explore automatic and adaptive methods for setting this parameter. Notice that all experiments required only a *single pass* through the data to compute answer likelihoods conditioned

	GPT-2				GPT-3			
	125M	350M	760M	1.6B	2.7B	6.7B	13B	175B
f_{\max}	8.48	14.78	13.88	14.29	17.33	19.42	22.06	26.76
Zhao et al. (2021)	17.45	22.87	23.90	23.97	26.30	30.57	31.96	34.78
CB ($\alpha_k = \alpha_k^*$)	19.85	22.87	25.74	25.43	28.75	32.25	35.02	37.57
α_k^*	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.4
k^*	1	2	3	3	1	1	1	2

Table 5: Accuracies (%) of GPT models on LAMA.

on full and premise-free contexts – no iterative gradient-based finetuning was applied.

Results: Knowledge retrieval. Unlike LAMBADA, where long contexts are required for inferring the last word, LAMA contains much shorter sentences for knowledge facts, i.e., (subject, relation, object). A recent study (Cao et al., 2021) shows that the prediction is biased by the relation in the short context, i.e., the answer to a prompt (e.g., “Dante was born in ___”) can be induced by the relation (“was born in”) without the subject. Coherence boosting mitigates the influence of those short contexts by making the prediction dependent on a longer context containing the subject.

We present results for all models on LAMA in Table 5. We also compare our model with contextual calibration (CC) (Zhao et al., 2021), which processes the LM’s output probabilities with a log-linear model.⁶ Coherence boosting with the selected α and k outperforms both the base model and CC by significant margins.

6 Extensions and future work

We suggest three promising research directions:

Coherence tuning. The need to evaluate the base LM with multiple contexts in coherence boosting introduces cost and complexity at inference time. It may be desirable instead to modify the weights of the base model to improve long-range coherence properties. In §B, we describe a ‘self-tuning’ algorithm that achieves this *without training on any data created for this purpose*.

New domains and architectures. In this paper, we mainly considered coherence boosting with decoder-only Transformer LMs trained on generic

text, but future work should consider other architectures and target domains. In §C, we give preliminary results on the *text summarization* domain.

Although we expect recency bias to be less pronounced in LMs that use separate attention modules to process the prompt and the output – such as encoder-decoder models for translation or summarization – procedures inspired by coherence boosting may prove effective in domains where a strong causal link between prompt and output is known to exist. Such domains include language generation conditioned on structured data (Yao et al., 2020; Mager et al., 2020; Moosavi et al., 2021) and model-guided reasoning in formal languages, such as proof or program synthesis (Polu and Sutskever, 2020; Chen et al., 2021; Li et al., 2022).

Efficient search proposals. Procedures that force LMs to be more focused on a prompt, or a specific part of it, when generating or ranking tokens can benefit algorithms that search for combinations of words through sampling. It would be interesting to use coherence boosting in non-autoregressive text generation algorithms, such as to accelerate the mixing of MCMC methods for constrained text generation (Miao et al., 2019; Zhang et al., 2020b; Malkin et al., 2021).

7 Conclusion

We have illustrated the hyposensitivity of pre-trained language models to long-range context and proposed a simple inference-time remedy. We hope to see coherence boosting used as a simple alternative or complement to finetuning procedures in zero-shot applications of pretrained LMs.

Acknowledgments

The authors are grateful to Sudha Rao, Matt Richardson, and Huan Sun for valuable discussions about this project. We thank the anonymous reviewers for their comments and suggestions.

⁶Note that CC applies a log-linear model to the *probability* domain, not the *logit* domain, which does not have an information-theoretic interpretation.

Ethics statement

We hope and expect to see a nonnegative net societal impact from better text generation and ranking algorithms in general and from this work in particular. As we have shown, there is room to improve the inference procedures used with small language models, which incur lower costs than training and evaluation of large models. However, researchers should bear in mind the risks and potential misuse of automatic generation of long-form text.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. *Neural Information Processing Systems (NeurIPS)*.
- Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint 2107.03374*.
- Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.
- George R. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology Research*.

- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28:337–407.
- Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and William B. Dolan. 2019. Grounded response generation task at DSTC7. *Dialog System Technology Challenges 7 (AAAI workshop)*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *International Conference on Learning Representations (ICLR)*.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. [Surface form competition: Why the highest probability answer isn’t always right](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. [Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- Alon Lavie and Abhaya Agarwal. 2007. [METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode.
- Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. [GPT-too: A language-model-first approach for AMR-to-text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online. Association for Computational Linguistics.
- Nikolay Malkin, Sameera Lanka, Pranav Goel, and Nebojsa Jojic. 2021. [Studying word order through iterative shuffling](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10351–10366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Clara Meister and Ryan Cotterell. 2021. [Language model evaluation beyond perplexity](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.
- Ning Miao, Hao Zhou, Lili Mou, Rui Yan, , and Li Lei. 2019. CGMH: Constrained sentence generation by Metropolis-Hastings sampling. *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct elec-](#)

- tricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.
- Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Learning to reason for text generation from scientific tables. *arXiv preprint arXiv:2104.08296*.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *arXiv preprint 2009.03393*.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations (ICLR)*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *AAAI Spring Symposium Series*.
- Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. *International Conference on Learning Representations (ICLR)*.
- Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on*

- Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. [Do long-range language models actually use long-range context?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems (NIPS)*.
- Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 200–207.
- Shaowei Yao, Tianming Wang, and Xiaojun Wan. 2020. [Heterogeneous graph transformer for graph-to-sequence learning](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7145–7154, Online. Association for Computational Linguistics.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Neural Information Processing Systems (NeurIPS)*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.
- Maosen Zhang, Nan Jiang, Lei Li, and Yexiang Xue. 2020b. [Language generation via combinatorial constraint satisfaction: A tree search enhanced Monte-Carlo approach](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1286–1298, Online. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Neural Information Processing Systems (NIPS)*.
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and William B. Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Neural Information Processing Systems (NeurIPS)*.
- Yizhe Zhang, Siqu Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning (ICML)*.
- Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *ACM SIGIR Conference on Research and Development in Information Retrieval*.

A On multi-objective training and log-linear weights

The section extends the discussion in §2.1.

Recall that the language model f is trained on the multi-objective loss (3):

$$\sum_{k=1}^M \lambda_k \underbrace{\mathbb{E}_{w_1 \dots w_{M+1} \in \mathcal{D}} [-\log f_k(w_{M+1} | w_1, \dots, w_M; \theta)]}_{\mathcal{L}_k(\theta)}, \quad \lambda_k = \frac{1}{M}.$$

As we saw in the main text, the scalarization weights λ_k are uniform as a consequence of the training regime. However, evaluation procedures effectively give nonuniform weight to the M prediction losses.

Some vector calculus. Denote by $\hat{\theta}(\lambda)$ a local optimum of the above optimization problem for general linear combination weights $\lambda = (\lambda_1, \dots, \lambda_M)$. Under suitable regularity conditions, the gradient of the combined loss vanishes:

$$\sum_k \lambda_k \left. \frac{\partial \mathcal{L}_k(\theta)}{\partial \theta} \right|_{\theta = \hat{\theta}(\lambda)} = \mathbf{0}. \quad (4)$$

Assuming the Hessian \mathbf{A} of the optimization criterion $\sum_k \lambda_k \mathcal{L}_k(\theta)$ is nonsingular, we can implicitly differentiate (4) with respect to λ to obtain the matrix derivative

$$\frac{\partial \hat{\theta}(\lambda)}{\partial \lambda} = -\mathbf{A}^{-1} \left. \frac{\partial (\mathcal{L}_1(\theta), \dots, \mathcal{L}_M(\theta))}{\partial \theta^T} \right|_{\theta = \hat{\theta}(\lambda)}. \quad (5)$$

The local dependence of the losses on the scalarization weights can be expressed as a bilinear form evaluated on $\frac{\partial \mathcal{L}_i}{\partial \theta}$ and $\frac{\partial \mathcal{L}_j}{\partial \theta}$:

$$\frac{\partial \mathcal{L}_i(\hat{\theta}(\lambda))}{\partial \lambda_j} = \left. \frac{\partial \mathcal{L}_i}{\partial \theta} \right|_{\theta = \hat{\theta}(\lambda)} \frac{\partial \hat{\theta}(\lambda)}{\partial \lambda_j} = -\frac{\partial \mathcal{L}_i}{\partial \theta} \mathbf{A}^{-1} \left. \frac{\partial \mathcal{L}_j}{\partial \theta^T} \right|_{\theta = \hat{\theta}(\lambda)}. \quad (6)$$

Because $\hat{\theta}$ is a local minimizer, $-\mathbf{A}^{-1}$ is negative definite. In particular, any $\frac{\partial \mathcal{L}_i(\hat{\theta}(\lambda))}{\partial \lambda_i}$ is negative. This expresses the intuitive fact that if an infinitesimally higher weight is given to some prediction loss in optimization, the value of this loss at the optimum will be infinitesimally lower.

For concreteness, consider how the highest-length prediction loss $\mathcal{L}_M(\hat{\theta}(\lambda))$ changes when λ_M is increased and the λ_j ($j \neq i$) are decreased with rate proportional to λ_j , while $\sum \lambda_j$ is kept constant. That is, let $\beta = (-\lambda_1, \dots, -\lambda_{i-1}, \sum_{j \neq i} \lambda_j, -\lambda_{i+1}, \dots, -\lambda_M)$. Then

$$\frac{d\mathcal{L}_i(\hat{\theta}(\lambda + t\beta))}{dt} = \sum_j \frac{\partial \mathcal{L}_i}{\partial \lambda_j} \beta_j = -\frac{\partial \mathcal{L}_i}{\partial \theta} \mathbf{A}^{-1} \sum_j \frac{\partial \mathcal{L}_j}{\partial \theta^T} \beta_j = -\frac{\partial \mathcal{L}_i}{\partial \theta} \mathbf{A}^{-1} \frac{\partial \mathcal{L}_i}{\partial \theta^T} \sum_j \lambda_j \leq 0, \quad (7)$$

where the last two equalities follow from (6) and (4), respectively, and the inequality holds because \mathbf{A}^{-1} is positive definite. So we have shown that, in nondegenerate cases, the $\mathcal{L}_M(\theta)$ term of the optimization criterion decreases under the locally optimal weights θ when λ_M is infinitesimally increased in this way.

Log-linear mixture of predictors. Returning to coherence boosting, suppose that we aim to build out of the predictors $f_k(-; \hat{\theta}(\lambda))$ a new predictor g that would have lower negative log-likelihood on prediction of a word given the maximum-length context:

$$\mathbb{E}_{w_1 \dots w_{M+1} \in \mathcal{D}} [-\log g(w_{M+1} | w_1, \dots, w_M)] < \mathbb{E} [-\log f_M(w_{M+1} | w_1, \dots, w_M; \hat{\theta}(\lambda))].$$

As we just saw, using this predictor in place of f_M achieves the same direction of movement in the prediction loss as optimizing with higher weight λ_M .

A naïve guess – not a proper predictor, as its outputs do not sum to 1 – would lightly perturb f_M by log-linearly mixing small multiples of the f_k weight weights β_k summing to 0:

$$g_{\text{naïve}}^{(t)}(w_1, \dots, w_M) = \exp \left(\log f_M(w_1, \dots, w_M; \hat{\theta}(\lambda)) + t \sum_k \beta_k \log f_k(-, \hat{\theta}(\lambda)) \right).$$

Then, by linearity of expectation,

$$\begin{aligned} \frac{d}{dt} \Big|_{t=0} \mathbb{E} \left[-\log g_{\text{naïve}}^{(t)}(w_{M+1} \mid w_1, \dots, w_M) \right] &= \sum_k \beta_k \mathbb{E} \left[-\log f_k(w_{M+1} \mid w_1, \dots, w_M; \hat{\theta}(\lambda)) \right] \\ &= \sum_k \beta_k \mathcal{L}_k(\hat{\theta}(\lambda)). \end{aligned} \quad (8)$$

This quantity is negative if, for example, $\mathcal{L}_M(\hat{\theta}(\lambda))$ is minimal among the $\mathcal{L}_k(\hat{\theta}(\lambda))$.

Reintroducing the normalization condition, we define a candidate function $g^{(t)}$ as the normalization of $g_{\text{naïve}}^{(t)}$ over w_{M+1} and compute, with the aid of (8) and using that the g_k are normalized to simplify the derivative of $\log \sum \exp$:

$$\begin{aligned} &\frac{d}{dt} \Big|_{t=0} \mathbb{E} \left[-\log g^{(t)}(w_{M+1} \mid w_1, \dots, w_M) \right] \\ &= \sum_k \beta_k \mathcal{L}_k(\hat{\theta}(\lambda)) + \frac{d}{dt} \Big|_{t=0} \mathbb{E} \log \sum_w g_{\text{naïve}}^{(t)}(w \mid w_1, \dots, w_M) \\ &= \sum_k \beta_k \mathcal{L}_k(\hat{\theta}(\lambda)) + \mathbb{E} \sum_w \left\langle \sum_k \beta_k \log f_k(w_1, \dots, w_M; \hat{\theta}(\lambda)), f_M(w_1, \dots, w_M; \hat{\theta}(\lambda)) \right\rangle \\ &= \sum_k \beta_k \mathcal{L}_k(\hat{\theta}(\lambda)) - \sum_k \beta_k \mathbb{E} \left[D_{\text{KL}}(f_M(w_1, \dots, w_M; \hat{\theta}(\lambda)) \parallel f_k(w_1, \dots, w_M; \hat{\theta}(\lambda))) \right], \end{aligned} \quad (9)$$

where the last line used that $\sum \beta_k = 0$.

In practice, we are interested in sparse log-linear mixtures. Taking $\beta_M = 1$, $\beta_k = -1$ for a single k , and all other $\beta_i = 0$, we conclude that the boosted model proportional to $f_M^{1+t} f_k^{-t}$ is a better predictor than f_M alone if the difference between prediction losses \mathcal{L}_M and \mathcal{L}_k is greater than the average KL divergence between the predictions f_M and f_k .

B From coherence boosting to coherence tuning

As mentioned in the main text, algorithms that modify the weights of a pretrained LM to increase effect of distant words, mimicking coherence boosting, are an interesting direction for future work. Here we propose an algorithm, **coherence tuning**, that achieves this without training on any specialized data.

Initializing with the pretrained model $f(-|-; \theta)$, the algorithm iterates the following training steps to bring the LM closer to its coherence-boosted version f_α :

- (1) Generate a sequence $w_1 \dots w_n$ from the current model $f(-|-; \theta)$.
- (2) Compute all next-token distributions under the coherence-boosted version of the current model ($f_\alpha(w_1 \dots w_k; \theta)$) and under the current model without boosting ($f(w_1 \dots w_k; \theta)$).
- (3) Gradient step on $\text{KL}(f_\alpha(w_1 \dots w_k; \theta) \parallel f(w_1 \dots w_k; \theta))$, where the first distribution f_α is treated as constant. This step may be restricted only to k near the end of the sequence.

We provide a batched implementation in Fig. B.1 in lieu of pseudocode. This coherence tuning code, which performs 32 gradient steps on batches of 32 sequences of length 32, runs in a few minutes on modern hardware, amortizing the overhead cost of coherence boosting while achieving comparable results on the WebText article completion task (second-to-last row of Table 2).

```

1  from transformers import GPT2LMHeadModel
2  import torch
3
4  model = GPT2LMHeadModel.from_pretrained('gpt2-large')
5  model.eval()
6
7  batch_size = 32
8  max_length, k = 32, 16
9  alpha = -0.1
10 num_steps = 32
11
12 opt = torch.optim.SGD(model.parameters(), 1e-2)
13
14 for _ in range(num_steps):
15     with torch.no_grad():
16         x = model.generate(None, max_length=max_length,
17                             early_stopping=False,
18                             do_sample=True,
19                             num_return_sequences=batch_size,
20                             pad_token_id=50256)
21
22         logits = model(x).logits[:, k:].log_softmax(2)
23         logits_short = model(x[:, :k]).logits
24         logits_cb = ((1-alpha)*logits + alpha*logits_short).log_softmax(2).detach()
25
26         kl = -((logits-logits_cb) * logits_cb.exp()).sum(2).mean()
27
28     opt.zero_grad()
29     kl.backward()
30     opt.step()

```

Figure B.1: Coherence tuning in PyTorch.

C GPT-2 summarization experiments

In §4 of the main text, we applied coherence boosting to generic text and dialogue response generation. Another interesting task that also requires long-range coherence is text summarization, in which the model is often expected to attend to the first few sentences to summarize a long article. Thus, we provide preliminary experiments for zero-shot abstractive summarization by applying our proposed method to GPT-2 models.

Experiment details. We take the two most popular summarization datasets, CNN/DM (See et al., 2017) and XSum (Narayan et al., 2018), where both contain recent articles and the summaries for the latter are more abstractive than the former. Following standard design (Radford et al., 2019), we append the tokens “TL;DR:” at the end of each article to induce summarization behavior of GPT models. We leverage the GPT-2 XL model and let it continue generating 100 tokens with greedy decoding. We take the first three sentences for CNN/DM articles and the first two sentences for XSum articles as their summaries. We use the preprocessed data and metric calculation from Zhong et al. (2020) and report the standard ROUGE scores in Table C.1.

To apply our proposed coherence boosting method, similarly to the method used for dialogue response generation, we define the short context as the newly generated text after the “TL;DR:” tokens. That is, at any time step during the summarization, the long context is the full article with the so-far generated summary, and the short context is only the generated summary.

Results. As we can see from Table C.1, our proposed CB method improves most of the metrics on both datasets. On the CNN/DM dataset, CB yields improvements of up to ~3 ROUGE points. We believe such a significant improvement is due to the article structure of the CNN/DM dataset. Specifically, the first three sentences in CNN/DM articles can provide pretty good summaries for a large portion of articles and have been considered as a very strong baseline for summarization models (Zhong et al., 2020). This

	CNN/DM			XSum		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
GPT-2 XL	26.671	7.792	23.926	21.346	4.360	16.880
CB, $\alpha = -0.1$	28.027	8.658	25.179	21.580	<u>4.265</u>	<u>17.025</u>
CB, $\alpha = -0.2$	28.995	9.293	26.066	<u>21.571</u>	4.200	17.026
CB, $\alpha = -0.3$	29.502	9.528	26.442	21.405	4.045	16.848
CB, $\alpha = -0.4$	29.772	9.663	<u>26.644</u>	21.150	3.876	16.613
CB, $\alpha = -0.5$	29.872	<u>9.625</u>	26.658	20.773	3.703	16.288
CB, $\alpha = -0.6$	<u>29.827</u>	<u>9.500</u>	26.524	20.379	3.525	16.010
CB, $\alpha = -0.7$	29.742	9.392	26.399	20.063	3.437	15.776
CB, $\alpha = -0.8$	29.703	9.304	26.242	19.661	3.305	15.460
CB, $\alpha = -0.9$	29.481	9.109	25.990	19.387	3.195	15.256

Table C.1: Abstractive summarization performance with the GPT-2 XL model. The best performance is bolded and the second-best is underlined.

long-range information in the first few sentences is hard to capture for GPT models, and this difficulty can be alleviated by coherence boosting.

Such structure might be less pronounced in the XSum dataset, thus leading to a smaller improvement from coherence boosting. The difference between the datasets is also reflected in the best value of α . The optimal α for the CNN/DM dataset is around -0.5 , while the best α for XSum is around -0.2 , which indicates that boosting has less effect on XSum.

Finally, we note that the performance with GPT models presented here is far lower than that of state-of-the-art summarization-specific models, such as finetuned BART models (Lewis et al., 2020; Zhang et al., 2020a). We leave the use of coherence boosting with such encoder-decoder models to future work.

D Example WebText completion

An example of human, top- p , and coherence boosting outputs is shown in Table D.1. All outputs for all boosting schemes are included in the code repository.

Prompt	University of Alberta doctoral candidate Scott Persons measured the limb length for 53 different species of carnivorous dinosaurs to create an equation to score how much each was built for speed.
Human	A new study published out of the University of Alberta has found that carnivorous dinosaurs evolved to become faster and faster over time — and the proof is in the leg length. As part of his doctoral research, lead author Scott Persons travelled the world measuring the limb lengths for 53 different species of carnivorous bipedal dinosaurs, including the Velociraptor, Allosaurus and Tyrannosaurus rex. In particular, Persons looked at the length of the dinosaurs’ legs below the knee. As a rule, he said, the longer the lower leg is in comparison to the upper leg, the faster the animal is. Modern-day cheetahs have proportionately longer legs than hyenas. That relationship is mirrored in the animals’ speeds, Persons said. "That’s true for modern carnivores, and must have been true for dinosaurs," Persons said in a statement released Wednesday.
Nucleus ($p = 0.95$)	"In a lot of areas of the world, dinosaurs at the time didn’t have legs, they had weapons," said Persons. So he used his scientific skill, titled "Toill’s Law of Tragedy," to give the dinosaurs an on-the-ground estimate of what their internal organs weighed. For example, wood-eating reptiles, such as the dinosaur Lufengosaurus and the ceratopsian Tyrannosaurus rex, would have weighed up to 52 kg (115 lbs) each. And as many as 50 plates should have been available to hunt, Persons said. But what’s important is that a dinosaur’s pack or herd probably wouldn’t have been able to feed all three in one sitting. Dinosaurs were extremely resilient, as very few of them survived from one kill to the next and later on would start a “phased trip” back to where they’d been killed, Persons said.
+ boost ($k = 64$, $\alpha_k = -0.1$)	"In a lot of areas of the world, dinosaurs at the time didn’t have legs, they had weapons," said Persons. So he used his undergraduate course to work on a formula to calculate a specific height for each carnivore based on the evolutionary requirements of different dinosaurs and when it would have been optimal for meat eating. He then moved that formula into his doctoral thesis project, which is looking at limb lengths for medium-sized, fearsome carnivores. "His thesis is about how to measure the shape of different places, plant diversity and evolutionary biology and kind of how we see dinosaurs and the relationships between species," said Susanne Dufout, an associate professor of paleontology at McMaster University who recently returned from attending Persons’ research. "It’s interesting because he’s basically done the legwork, developed this model that can give us information about plant species." ‘He was an amazing creature’

Table D.1: Completions of an article: written by a human (original WebText) and sampled from GPT-2 Large with top- p sampling, with and without coherence boosting. While top- p sampling produces text that is coherent at first glance – it is free of repetition and nonce words – the topic of the article meanders from limb length to internal organs and killing, and nonsensical comments appear (‘Toill’s Law of Tragedy’, herbivorous ceratopsian T-Rex, etc.). The output with coherence boosting is largely free of these issues, maintaining focus on limb length and diet.

E Prompt formats for multiple-choice tasks

Task	Prompt format
Story Cloze	<u>[Context]</u> <u>[Completion]</u>
HellaSwag	<u>[Context]</u> <u>he/she/they/...</u> <u>[Completion]</u>
COPA	<u>[Premise]</u> <u>because/so</u> <u>[Hypothesis]</u>
CommonsenseQA	<u>[Question]</u> <u>the answer is:</u> <u>[Answer]</u>
OpenBookQA	<u>[Question]</u> <u>the answer is:</u> <u>[Answer]</u>
ARC Easy	Question: <u>[Question]</u> Answer: <u>[Answer]</u>
ARC Challenge	Question: <u>[Question]</u> Answer: <u>[Answer]</u>
PIQA	Question: <u>[Question]</u> Answer: <u>[Answer]</u>
SST-2	<u>[Context]</u> <u>This quote has a tone that is:</u> <u>[Label]</u>
SST-5	<u>[Context]</u> <u>This quote has a tone that is:</u> <u>[Label]</u>
AGNews	Title: <u>[Title]</u> Summary: <u>[Context]</u> Topic: <u>[Label]</u>
TREC	<u>[Question]</u> <u>The answer to this question will be</u> <u>[Label]</u>
BoolQ	<u>[Passage]</u> \n Question: <u>[Hypothesis]</u> <u>True or False?</u> Answer: <u>[Label]</u>
RTE	<u>[Premise]</u> \n question: <u>[Hypothesis]</u> <u>true or false?</u> \n answer: <u>[Label]</u>
CB	<u>Given question: [Premise]</u> <u>Is [Hypothesis] true, false or neither?</u> \n <u>The answer is: [Label]</u>

Table E.1: Prompt formats used in our experiments. The full context is underlined in blue; the premise-free context is also underlined in red. We mainly draw inspiration from (Brown et al., 2020; Holtzman et al., 2021; Zhao et al., 2021) to make our prompts more natural to facilitate boosting the coherence of the completion.

F Additional results

	GPT-3 Small				GPT-3 Medium				GPT-3 Large				GPT-3 XL			
	f_{\max}	$\alpha = 1$	$\alpha = \alpha^*$	α^*	f_{\max}	$\alpha = 1$	$\alpha = \alpha^*$	α^*	f_{\max}	$\alpha = 1$	$\alpha = \alpha^*$	α^*	f_{\max}	$\alpha = 1$	$\alpha = \alpha^*$	α^*
Story Cloze	66.0	70.9	74.5	-0.8	70.1	76.3	78.0	-0.8	74.2	82.9	80.8	-0.7	79.3	82.9	86.9	-0.6
HellaSwag	35.7	38.9	42.0	-0.9	42.8	46.8	51.3	-0.8	50.5	55.1	62.2	-0.8	59.2	62.7	72.3	-0.8
COPA	73.0	71.0	75.0	-0.6	85.0	79.0	83.0	-0.7	84.0	83.0	84.0	-0.6	93.0	87.0	94.0	-0.5
CsQA	34.6	46.4	48.0	-0.7	42.4	51.4	53.0	-0.7	50.0	57.5	60.4	-0.7	61.1	68.0	70.4	-0.7
OBQA	16.0	39.8	46.6	-2.2	16.4	41.8	48.8	-1.4	20.8	45.4	47.8	-1.6	28.0	52.2	52.6	-1.1
ARC-E	51.3	48.1	56.0	-0.5	59.8	54.8	63.3	-0.4	68.4	60.3	70.7	-0.5	76.2	69.2	78.3	-0.4
ARC-C	22.6	30.8	31.1	-1.4	27.5	35.3	35.5	-1.2	33.9	41.8	41.8	-0.9	43.9	50.6	49.2	-1.1
PIQA	69.0	57.5	69.6	-0.4	74.4	60.4	74.7	-0.4	76.3	64.2	77.7	-0.4	79.3	66.3	78.9	-0.6
SST-2	70.6	79.8	84.6	-2.3	69.5	75.2	88.0	-4.8	66.8	65.2	70.0	2.0	86.2	88.1	89.8	-0.5
SST-5	26.7	26.6	26.1	-1.1	29.3	30.7	30.0	-1.2	28.1	33.2	30.1	-0.8	31.2	34.8	38.5	-1.4
AGNews	67.1	69.2	69.5	-1.2	63.3	64.8	65.4	-2.0	69.2	65.7	69.5	-0.3	71.7	71.7	71.8	0.2
TREC	28.8	57.2	57.4	-1.0	30.2	62.6	63.6	-0.8	35.2	28.8	37.2	-0.3	52.4	47.0	56.0	-0.6
BoolQ	60.7	62.4	62.2	-1.4	61.6	63.4	63.5	-0.9	64.2	65.6	68.1	-4.5	71.6	73.7	72.7	-0.4
RTE	49.8	51.3	51.3	-3.6	54.5	50.5	49.1	-1.2	53.8	55.6	55.2	-1.4	56.0	57.4	60.3	-0.6
CB	33.9	19.6	21.4	-0.7	8.9	25.0	39.3	-1.9	32.1	28.6	32.1	-0.2	5.4	25.0	28.6	-1.9
average	47.1	51.3	54.4	-1.3	49.0	54.5	59.1	-1.3	53.8	55.5	59.2	-0.8	59.6	62.4	66.7	-0.7

Table F.1: Accuracy (%) of GPT-3 models on all multiple-choice tasks, in the same format as Table 4.

	GPT-2 Small				GPT-2 Medium				GPT-2 Large				GPT-2 XL			
	f_{\max}	$\alpha = -1$	Ours	α^*	f_{\max}	$\alpha = -1$	Ours	α^*	f_{\max}	$\alpha = -1$	Ours	α^*	f_{\max}	$\alpha = -1$	Ours	α^*
Story Cloze	59.9	64.8	64.2	-1.0	63.0	68.5	70.4	-0.7	66.0	72.0	74.4	-0.8	67.6	75.1	76.8	-0.7
HellaSwag	28.9	31.0	31.8	-0.9	33.4	36.6	38.1	-0.9	36.6	39.5	43.0	-0.8	40.0	42.6	47.7	-0.8
COPA	62.0	56.0	64.0	-0.7	69.0	69.0	72.0	-0.6	69.0	60.0	69.0	-0.6	73.0	70.0	77.0	-0.4
CsQA	29.5	42.3	43.2	-0.8	31.3	44.6	45.3	-0.8	35.7	47.3	50.0	-0.8	37.8	50.5	52.9	-0.8
OBQA	11.2	30.6	40.8	-1.6	15.6	34.8	43.8	-2.1	13.6	34.4	44.2	-1.8	15.6	38.4	47.0	-1.9
ARC-E	43.8	42.1	46.0	-0.3	49.1	44.5	51.3	-0.6	53.2	46.5	56.2	-0.5	58.3	51.4	60.3	-0.4
ARC-C	19.0	26.1	29.1	-4.2	21.5	27.3	27.0	-1.0	21.7	28.3	29.1	-2.8	25.0	33.5	34.4	-1.1
PIQA	62.9	57.5	63.4	-0.6	67.6	56.1	68.1	-0.5	70.3	60.0	70.1	-0.4	70.8	60.4	71.5	-0.4
SST-2	65.7	74.7	82.3	-2.2	72.6	83.5	88.2	-2.0	77.2	87.6	88.0	-1.2	86.4	84.5	86.9	-0.1
SST-5	25.9	30.9	30.9	-1.2	20.5	33.3	35.2	-1.1	29.1	31.8	35.2	-1.4	28.7	38.7	36.9	-1.7
AGNews	58.6	60.8	62.2	-0.6	64.6	66.5	66.3	-0.7	62.6	62.1	63.8	-0.4	67.2	67.4	68.3	-0.4
TREC	23.4	29.6	32.2	-0.8	27.4	17.6	36.0	-0.4	22.6	45.4	44.2	-1.2	23.4	27.4	40.0	-0.8
BoolQ	49.4	58.1	62.1	-3.0	56.6	61.8	61.8	-0.9	61.2	62.3	62.2	-1.8	62.1	63.5	63.2	-0.6
RTE	51.3	49.8	53.4	-0.3	53.1	50.9	53.8	-0.2	53.1	46.6	50.2	-1.2	49.1	48.7	49.1	0.9
CB	12.5	23.2	48.2	-2.4	8.9	37.5	55.4	-2.5	8.9	32.1	53.6	-2.5	30.4	51.8	66.1	-1.9
average	40.3	45.2	50.3	-1.4	43.6	48.8	54.2	-1.0	45.4	50.4	55.5	-1.2	49.0	53.6	58.5	-0.7

Table F.2: Accuracy (%) of GPT-2 models on all multiple-choice tasks, in the same format as Table 4.

	GPT-3 Small			GPT-3 Medium		GPT-3 Large		GPT-3 XL		
	PMI	CC	Ours	PMI	Ours	PMI	Ours	PMI	CC	Ours
Story Cloze	73.1	-	74.5	76.8	78.0	79.9	80.8	84.0	-	86.9
HellaSwag	34.2	-	42.0	40.0	51.3	45.8	62.2	53.5	-	72.3
COPA	74.4	-	75.0	77.0	83.0	84.2	84.0	89.2	-	94.0
CsQA	44.7	-	48.0	50.3	53.0	58.5	60.4	66.7	-	70.4
OBQA	42.8	-	46.6	48.0	48.8	50.4	47.8	58.0	-	52.6
ARC-E	44.7	-	56.0	51.5	63.3	57.7	70.7	63.3	-	78.3
ARC-C	30.5	-	31.1	33.0	35.5	38.5	41.8	45.5	-	49.2
SST-2	72.3	71.4	84.6	80.0	88.0	81.0	70.0	71.4	75.8	89.8
SST-5	23.5	-	26.1	32.0	30.0	19.1	30.1	29.6	-	38.5
AGNews	67.9	63.2	69.5	57.4	65.4	70.3	69.5	74.7	73.9	71.8
TREC	57.2	38.8	57.4	61.6	63.6	32.4	37.2	58.4	57.4	56.0
BoolQ	53.5	-	62.2	61.0	63.5	60.3	68.1	64.0	-	72.7
RTE	51.6	49.5	51.3	48.7	49.1	54.9	55.2	64.3	57.8	60.3
CB	57.1	50.0	21.4	39.3	39.3	50.0	32.1	50.0	48.2	28.6

Table F.3: Performance comparison with other inference methods on GPT-3 models. PMI (Holtzman et al., 2021) is an unconditional probability normalization method, CC (Zhao et al., 2021) is the contextual calibration method. We compare them in the zero-shot setting.

	GPT-2 Small		GPT-2 Medium		GPT-2 Large			GPT-2 XL		
	PMI	Ours	PMI	Ours	PMI	Channel	Ours	PMI	CC	Ours
Story Cloze	67.0	64.2	71.6	70.4	73.4	-	74.4	76.3	-	76.8
HellaSwag	29.1	31.8	32.8	38.1	35.1	-	43.0	37.8	-	47.7
COPA	62.8	64.0	70.0	72.0	69.4	-	69.0	71.6	-	77.0
CsQA	36.4	43.2	41.8	45.3	44.5	-	50.0	47.8	-	52.9
OBQA	32.4	40.8	38.6	43.8	43.2	-	44.2	46.0	-	47.0
ARC-E	39.3	46.0	42.4	51.3	47.0	-	56.2	49.9	-	60.3
ARC-C	28.2	29.1	28.6	27.0	31.6	-	29.1	33.8	-	34.4
SST-2	67.1	82.3	86.2	88.2	85.6	77.1	88.0	87.5	82.0	86.9
SST-5	30.0	30.9	39.3	35.2	22.0	29.2	35.2	40.8	-	36.9
AGNews	63.0	62.2	64.4	66.3	64.1	61.8	63.8	65.4	60.0	68.3
TREC	36.4	32.2	21.6	36.0	44.0	30.5	44.2	32.8	37.3	40.0
BoolQ	51.1	62.1	49.7	61.8	46.7	-	62.2	49.5	-	63.2
RTE	49.8	53.4	54.9	53.8	54.2	-	50.2	53.4	48.5	49.1
CB	50.0	48.2	50.0	55.4	50.0	-	53.6	50.0	17.9	66.1

Table F.4: Performance comparison with other inference methods on GPT-2 models. PMI (Holtzman et al., 2021) is an unconditional probability normalization method, CC (Zhao et al., 2021) is the contextual calibration method and Channel (Min et al., 2021) uses an inverted-LM scoring approach that computes the conditional probability of the input given the label. We compare them in the zero-shot setting.

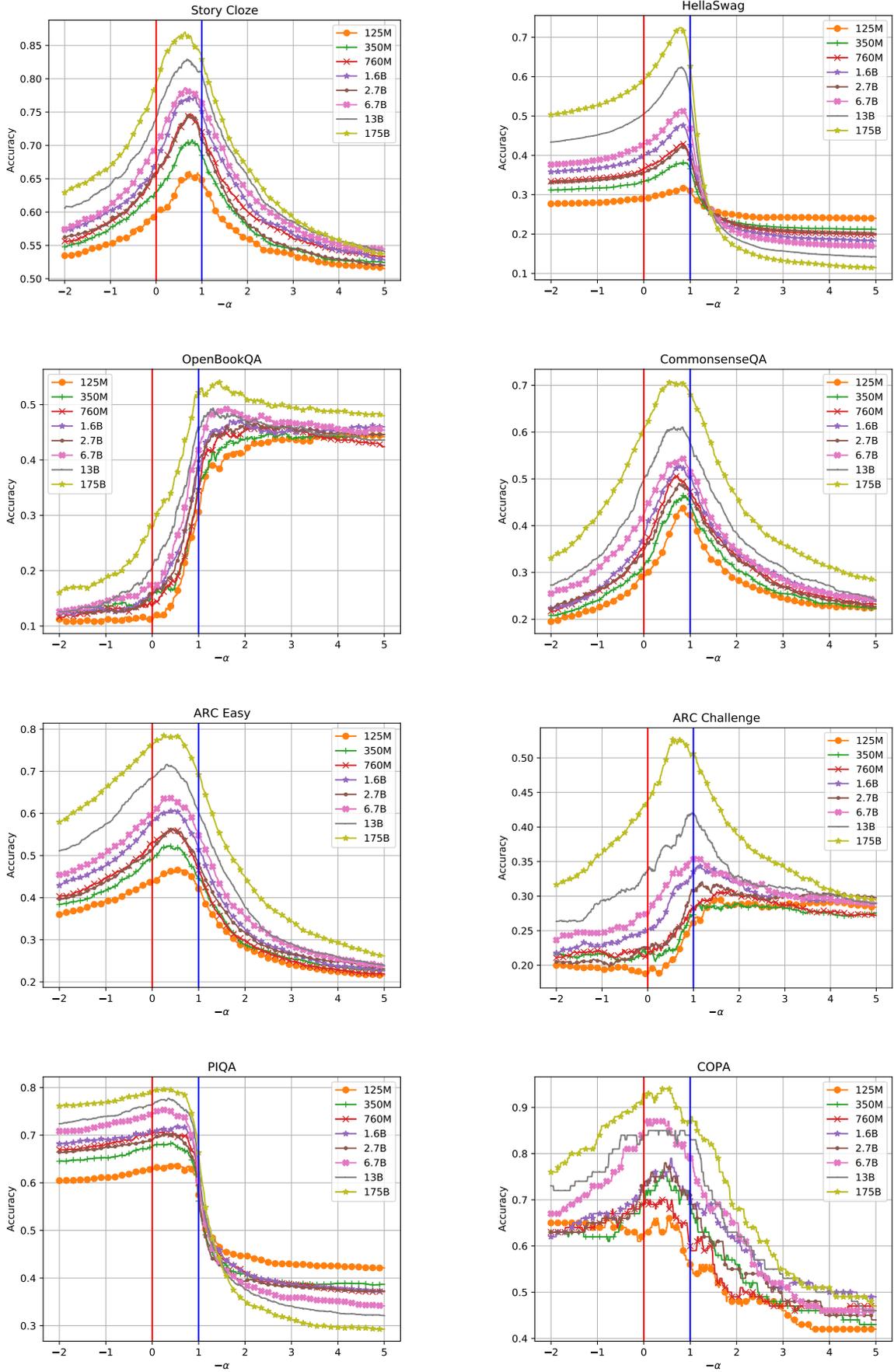


Figure F.1: Model comparison for StoryCloze, HellaSwag, OpenBookQA, CommonsenseQA, ARC Easy, ARC Challenge, PIQA and COPA by varying α on the testing set.

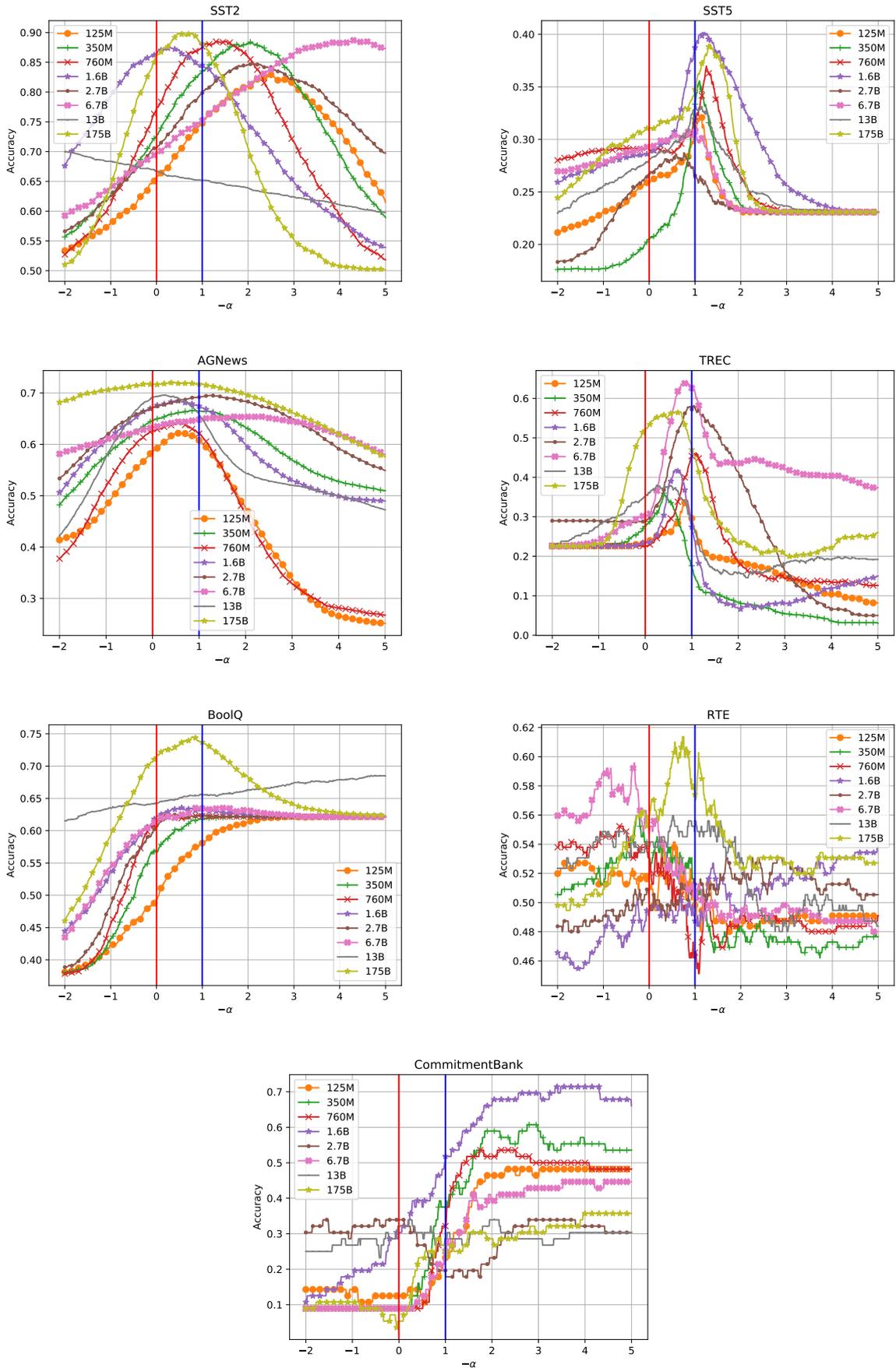


Figure F.2: Model comparison for SST-2, SST-5, AGNews, TREC, BoolQ, RTE and CommitmentBank by varying α on the testing set.