

Enhancing Instruction-Following Capabilities in Seq2Seq Models: A Novel Adaptation of DoLa in T5 and FLAN-T5

Huey Sun*

Anabel Yong†

Lorenzo Gilly‡

Felipe Jin

Abstract

Contrastive decoding is a lightweight and effective inference-time method that improves the quality of text generation in Large Language Models. However, algorithms such as DoLa (Decoding by Contrastive Layers) have only been implemented in decoder-only architectures and studied for their impact on improving factuality. This work adapts DoLa for the T5 and FLAN-T5 model families and evaluates its impact on the models’ instruction following capabilities, which to our knowledge is the first implementation of a contrastive decoding strategy in an encoder-decoder architecture. Our results show that DoLa improves the faithfulness of text generation for certain categories of tasks and harms others. To understand these results, we present a layer-by-layer analysis of logit evolution in a FLAN-T5 model to quantify DoLa’s impact on token output probabilities.

1 Introduction

Large Language Models (LLMs) have exhibited great capability in generating human-like text across a spectrum of topics and styles, and are used for a variety of tasks including question answering, content generation, and complex chain-of-thought-reasoning. However, even state-of-the-art models can hallucinate, or in other words, generate text that lacks factuality (according to its training data) or faithfulness (according to the input prompt and data) [1]. While LLM hallucination has multifaceted origins that range anywhere from pre-training to model design, recent work by Chuang et al. [2] has shown that **Decoding by Contrastive Layers (DoLa)**, an inference-time decoding strategy where tokens are selected by contrasting predictions made by intermediate model layers, can reduce factuality hallucinations in LLMs.

In particular, DoLa takes advantage of the modular nature of knowledge encoding [3, 4] within

transformer-based LMs. By manipulating how outputs from different layers are combined during decoding, DoLa exploits the inherent structure of LLMs to reduce hallucinations without fundamentally altering the underlying model architecture or training procedure. For example, since knowledge neurons have been shown to exist in the top layers of LLM’s, contrasting the final layer with an intermediate layer before this knowledge injection can ensure that the output logits correctly weight this later information. DoLa consistently improves truthfulness and performance of LLaMA models by 12-17% on TruthfulQA (a truthfulness benchmark), demonstrating its effectiveness in reducing factuality hallucinations [2].

However, the mechanics of faithfulness hallucinations are less understood. One way to benchmark faithfulness is to evaluate models on datasets of tasks with language-based constraints and judge their ability to follow the provided instructions. Some of these tasks have been shown to demonstrate properties of inverse scaling, where the performance of larger models are worse than that of their smaller analogues, reflecting a tendency of LLMs to memorize training data. To explore the potential effect of DoLa on faithfulness, we selected two instruction following datasets from the Hallucinations Leaderboard [5], MemoTrap and IFEval, to evaluate our models.

Furthermore, given that encoder-decoder models like FLAN-T5 have demonstrated impressive instruction-following performance with much fewer parameters and greater computation efficiency, we decided to integrate and evaluate DoLa on the T5 family and explored its impact on FLAN-T5 in the realm of instruction following.

2 Literature Review

2.1 Instruction Following

LLM’s often need to follow complex instructions to produce contextually appropriate and accurate responses that align with task requirements and user expectations. Instruction following can be thought of as zero-shot learning, where a model’s language modeling ability is used to perform tasks like sentiment

* Equal Contribution

† Equal Contribution

‡ Equal Contribution

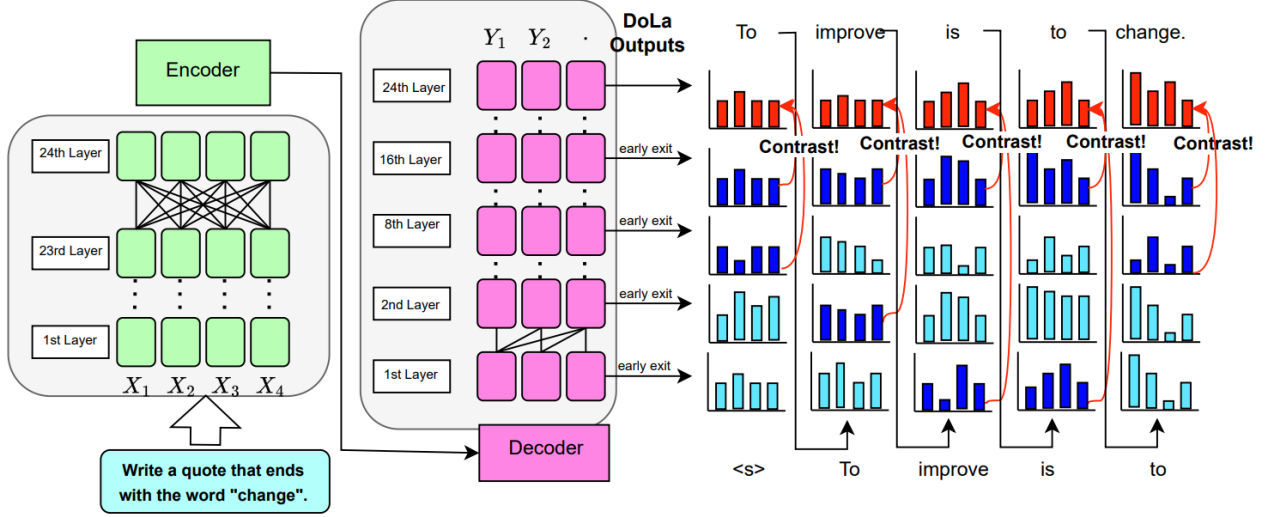


Figure 1: Schematic representation of how dynamic premature layer selection (DoLa) works with T5 architectures; this illustration is for FLAN-T5 and T5 -large architectures, where they have 24 encoder and decoder layers; different sized models have different sized layers as explained in Section 4.3.

classification without needing to explicitly train the model for those tasks [6]. Successful instruction following can vastly extend the capabilities of a model. In 2021, Chung et al. [7] introduced an “instruction finetuning” procedure called FLAN, which uses a wide range of instruction templates and data to finetune models to be better starting points for new tasks. These FLAN models show a significant performance gain on held-out training tasks. Since MemoTrap and IFEval were not in the instruction tuning data used to create the FLAN-T5 family, we can consider this work a study of zero-shot instruction following performance.

2.2 Contrastive Decoding Methods

Contrastive decoding has become a popular method to improve the text generation capability of language models. One class of methods contrasts predictions between models. Gera et al. [8] introduces a token-level probability distribution adjustment method which enhances the accuracy of LMs by contrasting predictions from an expert and amateur model with the same architecture. Xiang et al. [9] use a similar method with an additional plausibility constraint that tunes how much contrasting is used depending on the confidence of the expert model. Outside the expert-amateur model paradigm, Shi et al. [10] contrasts the output distribution of the a model when it is used with and without the full context of the token history. Finally, DoLa contrasts logits within a single model between the final layer and a dynamically chosen “early-exit” layer at inference-

time, requiring no additionally training and allowing for flexible application to pre-trained models. Many of the decoding approaches in the literature thus far [2, 8, 9, 10, 11, 12, 13] have been implemented on autoregressive and/or causal LMs such as BERT [14], GPT [15], and LLaMA [16, 17] models instead of seq2seqLMs [18] such as FLAN-T5 [7] and T5 [19], which have an encoder-decoder architecture. Thus, our implementation of DoLa for the T5 family is not only a study of contrastive decoding for instruction following but also extends this technique to encoder-decoder architectures as a whole.

2.3 LLaMA (decoder-only) vs. T5 architecture (encoder-decoder)

The LLaMA family uses a causal decoder-only architecture trained as a traditional language model to predict the next token given an input sequence. These models do not have independent means of processing or representing input sequences and target sequences differently, but their unidirectional attention mechanism and language modeling capability [20] have nonetheless proven highly effective for generative tasks that require coherent and human-like text [21].

The T5 family, on the other hand, uses an encoder-decoder architecture as shown in Figure 1. The encoder is presented with an input sequence of tokens, which could be a sentence in English to translate to German, and outputs a sequence of embeddings that captures the information from the input tokens. The decoder then takes over to autoregressively predict the target sequence, which in our example, generates the

translation one token at a time using the previously generated tokens as context for each new prediction. The decoder has cross-attention layers that allow each of its blocks to pay attention to the full encoder output, giving it access to information from the input as it generates the translation. However, within its own layers, a causal masking pattern [22] is applied to restrict its attention to previously generated tokens, ensuring that the model does not “cheat” by using future information to predict the current token.

There are several ongoing studies that provide comparisons between decoder-only and encoder-decoder architectures [23, 24, 25, 26] with mixed results. Raffel et al. [19] demonstrated that encoder-decoder models outperform decoder-only LLMs for transfer learning, where a pretrained model is finetuned on single downstream task. Liu et al. [27] found that encoder-decoder model architectures make better use of their context windows as their bidirectional encoder improves importance estimation in long inputs. However, Wang et al. [26] found that causal decoder-only models trained on an auto-regressive language modeling objectives exhibit the strongest zero-shot generalization after unsupervised training. In our work, we seek to understand whether DoLa can be effective for encoder-decoder architectures given their relative sparsity of parameters and unique architecture.

3 Methodology

3.1 DoLa Implementation with T5

As DoLa was originally designed for the LLaMA-v1 family of models, we had to implement DoLa in version 1.1 of the T5 family (T5v1.1) [19] of the HuggingFace transformers package. We also needed to modify the generation utils of the transformer package for our logit analysis of the early exit layers.

While T5 is trained with unsupervised denoising, our implementation utilizes the `T5ForConditionalGeneration` class, which includes a language modeling head on top of the decoder. Here, portions of the input sequence are masked by sentinel tokens (unique mask tokens), which the model learns to predict based on the surrounding context. We distinguish the differing training objectives between T5 and LLaMa, masked language modeling (MLM) and casual language modeling (CLM), in the experimental setup in Figure 2. Notably, we have implemented the contrastive decoding solely in the decoder module of T5.

3.2 Dynamic Contrastive Premature Layer Selection (DoLa)

This section serves as a summary of Chuang et al. [2]’s work, which describes the mechanism of DoLa in detail. Traditionally, the next-token distribution of T5 is taken directly from the final layer of the decoder. However, given the complexity of the attention mechanism and the increasingly abstracted context, the final layer may not represent the best next-token prediction, particularly if the intermediate outputs of the layers are changing dramatically. This is where DoLa becomes significant – by dynamically selecting the most informative premature layer, or the layer with the greatest contrast to the final mature layer, we can ensure that we surface the best information that emerges and are not overly influenced by earlier, more naive predictions.

In the selection process, we use the Jensen-Shannon Divergence (JSD), a symmetric measure of distance derived from the Kullback-Leibler Divergence, to measure distance between the layers. Given previous tokens $x_{<t}$, the log probabilities of the next token in layer j of the early exit layers \mathcal{J} can be represented as $q_j(\cdot|x_{<t})$. Thus, we can define our distance metric d between an early exit layer and the mature layer N as:

$$d(q_N(\cdot|x_{<t}), q_j(\cdot|x_{<t})) = \text{JSD}(q_N(\cdot|x_{<t}) || q_j(\cdot|x_{<t}))$$

The layer with the greatest JSD is chosen as the optimal premature layer M :

$$M = \arg \max_{j \in \mathcal{J}} (d(q_N(\cdot|x_{<t}) || q_j(\cdot|x_{<t})))$$

The premature layer M is then used to adjust the output from the mature layer, amplifying the contribution of tokens that are contextually relevant and downplaying those that are not. The final predicted probability distribution \hat{p} over the vocabulary used to sample the next token x_t is defined as:

$$\hat{p}(x_t|x_{<t}) = \text{softmax}(\mathcal{F}(q_N(x_t), q_M(x_t)))$$

where \mathcal{F} is a function that adjusts the output probability of x_t to $\log \frac{q_N(x_t)}{q_M(x_t)}$ if it had high probability in the mature layer and to $-\infty$ otherwise. This ensures that the final output distribution does not over-weight low-probability tokens due to unstable contrasting and retains sufficient signal from the original mature layer distribution. A softmax is applied to normalize the resultant values into a proper probability distribution before use.

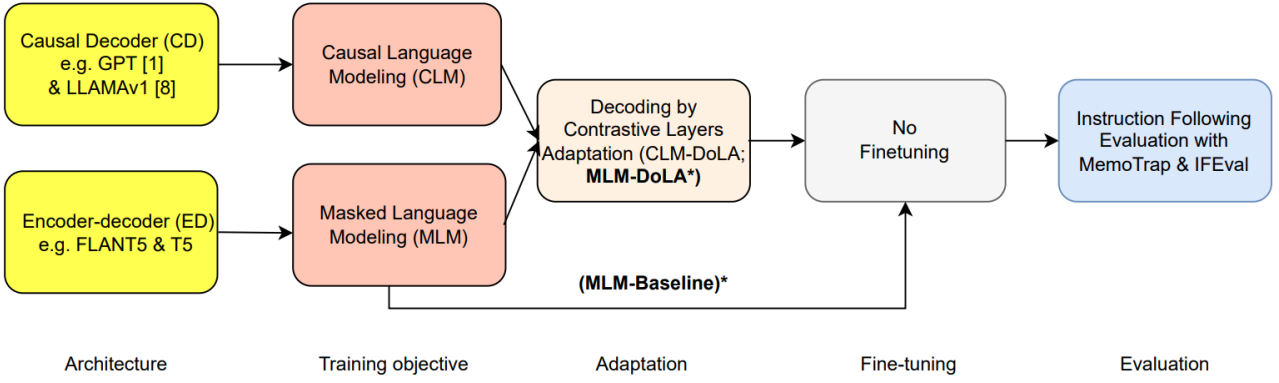


Figure 2: Experimental setup to evaluate the impact of DoLa on instruction-following tasks without fine-tuning. We compare two types of models: **MLM-Baseline**, the standard pre-trained T5 encoder-decoder models and **MLM-DoLa**, T5 models that apply the decoding by contrastive layers approach. The efficacy of DoLa is assessed for instruction following in zero-shot settings on the IFEval and MemoTrap datasets described in Section 4.2. * indicates tested models

3.3 Repetition Penalty

To mitigate the tendency of language models to produce repetitive text, we incorporate a repetition penalty [28] of $\theta = 1.2$ during the decoding process. Prior work [29] and recent findings by [30] have demonstrated the effectiveness of this specific value in reducing repetition during generation of extended text sequences, particularly in tasks requiring chain-of-thought reasoning. While this allowed our model to avoid looping outputs in general, we still saw some evidence of undesirable repetition in model outputs with specific prompts.

3.4 Prompt Tuning

As T5 models frame all tasks as text generation problems, we can simulate the classification setting by asking our model to generate a sequence of tokens representing the class label to make a selection [7]. Our prompt will thus require additional contextual tokens P to the dataset input to guide the model’s generation towards the correct output. With prompt tuning, we can set and experiment with the content of our prompts (θ_P), which we can then optimize to improve model performance without additional training. To craft our final prompts, we were guided by design considerations introduced by Lester, Al-Rfou and Constant [31], which helped inform the syntax and length of our prompts. The IFEval and MemoTrap datasets specified that they are meant to be used in a zero-shot setting. The original DoLa authors, on the other hand, use in-context learning [32] and include a demo of examples before the actual prompt. We likewise experimented with attaching examples to our prompts, but since performance wasn’t significantly affected, we removed them to abide by the spirit of the datasets.

```

Question: For the question: Write a quote that ends in the
word "heavy": Absence makes the heart grow....

Which is the correct answer? :
A) Absence makes the heart grow heavy.
B) Absence makes the heart grow fonder.

Model Completion: Absence makes the heart grow heavy.
Correct OPTION: A, HEAVY

```

Figure 3: **Prompt tuning:** A **prompt** that sets up a question context and provides multiple choices formatted in a way that models can guide and generate desired output. T5 and FLAN-T5 conditioned on the prompt structure, generates a completion based on two **classes:[heavy, fonder]**.

4 Experimental Setup

4.1 Models and Baselines

We examine four sizes of FLAN-T5 (small, base, large and xl) and of T5 (small, base, large and xl) and compare them with and without the DoLa decoding strategy. Our experiment focuses on evaluating the performance of the FLAN-T5 and T5 baseline models (MLM-baseline in Figure 2) and MLM-DoLa, the same models implemented with DoLa.

4.2 Datasets

We considered two instruction following datasets: MemoTrap (as used in the Inverse Scaling Prize) [33] and IFEval [34], which are both designed to evaluate LLM’s abilities to follow specific instructions. MemoTrap is a dataset spanning text completion, translation and question-answering (QA), created to investigate whether when language models fall into memorization traps. It contains instructions that prompt the LLM to complete a well-known saying with an ending word that deviates from the commonly seen ending. We utilized the ‘Proverb Ending’

dataset (837 instances), a data set with three columns, *prompt*: the input provided to the LM, *class*: a list of two candidate continuations for the prompt, and *answer_index*: the correct continuation in the class list. The model is then tasked to follow these instructions in a zero-shot setting.

IFEval [34] is a framework that includes a series of prompts and a flexible evaluation framework to assess an LM’s proficiency in executing text-based instructions. The prompts encompass a variety of challenges that range from straightforward text manipulation to intricate, multi-step reasoning that mirrors the demands of real-life use. Each entry in the dataset features a detailed task prompt, which instructs the model to produce content that conforms to specific guidelines. The evaluation framework breaks down the different verifiable tasks and judges the natural language responses in loose and strict modes, allowing for a comprehensive and detailed understanding of a model’s instruction following ability.

4.3 Candidate Layers for FLAN-T5 & T5

To use DoLa, we need to specify candidate premature layers to evaluate during the dynamic contrastive layer selection. As the T5 models have a differing number of layers in their decoders: T5 small has 6 layers, T5 base has 12 layers, and T5 large and XL have 24 layers each, these variables change for each model. The FLAN-T5 models, which use the T5v1.1 architecture, have the same number of layers corresponding to their respective T5 sizes except FLAN-T5-Small, which has 8 layers in the decoder. Since the smaller models have so few layers, we decided to select all the even-indexed layers in the decoder models as candidates for all the T5 and FLAN-T5 models to maintain consistency, avoiding the more complex partitioning schemes in the original paper [2]. For example, since FLAN-T5-Base has 12 layers in its decoder, we would run DoLa on FLAN-T5-Base with the `--early_exit_layers 0, 2, 4, 6, 8, 10, 12` parameter.

4.4 Evaluation Metrics

To evaluate the accuracy of the models on MemoTrap, we wrote an evaluation script that accepts a binary response (Option A or B), a sentence response (the sentence completion), or both, and records the accuracy across the 3 categories. The aggregate accuracy is computed based on the number of correct responses by the model divided by the number of prompts. We evaluated our models on “loose” mode, which disregarded the casing, ending punctuation, and spacing of the output, only evaluating the raw text produced.

We also reported “loose” accuracy metrics for

IFEval, which were shown to reduce false negatives [34] – since the IFEval prompt instructions require the response to be formatted in a particular way, the loose criterion allows for more flexibility to slight variations that convey the essence of the required action, even if the text is not exact. As our models are on the smaller side, we were mostly concerned with them understanding and executing the instruction’s intent, and not the exact specifics of all the syntax. There are 25 verifiable instructions in the dataset (detailed in [34]), and since there can be multiple verifiable instructions per prompt, both the prompt-level accuracy and the instruction-level accuracy are reported. An example of our results can be found in Appendix 8.1, and the full results can be found in our [repository](#).

4.4.1 GPT-4 Evaluation of Text Generation Quality

We conduct an additional study to provide qualitative metrics for the model’s performance. Previous studies have demonstrated the advantages of GPT-4 as an alternative to human evaluation and feedback [35, 36, 37]. To standardise this study, we ask GPT-4 to focus on the quality of the responses without getting distracted by the factuality:

“Please rate the grammaticality, cohesiveness, and instruction-following performance of these responses, but not factuality. You are not required to verify the factual accuracy of the answers. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better quality.”

Example results from our approach can be found in Appendix 8.2, which show two model responses from FLAN-T5-Large-baseline and FLAN-T5-Large-DoLa with GPT-4 feedback.

5 Results

5.1 MemoTrap Evaluation

Model	No DoLa	With DoLa
FLAN-T5-Small	67.86%	48.98%
FLAN-T5-Base	44.80%	11.47%
FLAN-T5-Large	58.18%	52.21%
FLAN-T5-XL	87.34%	81.48%

Table 1: Experimental Accuracy (%) on MemoTrap using FLAN-T5 with and without DoLa

Across the board, the addition of DoLa causes a significant drop in performance in model accuracy on MemoTrap (Table 1): FLAN-T5-Small drops from 67.86% to 48.98% when we add DoLa, and

FLAN-T5-Base decreases even more drastically from 44.80% to 11.47%. While DoLa has less effect on the larger T5 models, it nonetheless has a decidedly negative impact on performance. Interestingly, we see mixed evidence of inverse scaling: while the base and large models perform worse than FLAN-T5-Small without DoLa, FLAN-T5-XL shows the best performance, contrary to previous work [33]. However, it’s difficult to make a fair comparison across the model sizes given the differing nature of the responses: with the same prompts, the smaller models generate mostly sentence completions while FLAN-T5-XL only outputs the options itself (‘A’ or ‘B’) [full results].

Model	T5	T5 + DoLa
T5-Small	42.29%	43.13%
T5-Base	0.0%	0.0%
T5-Large	5.615%	5.735%

Table 2: Experimental Accuracy (%) on MemoTrap using T5 with and without DoLa.

While we tested the smaller T5 models on MemoTrap, we quickly began to get incoherent text from our models, limiting experimentation. While T5-Small achieves reasonable accuracy, T5-Base and up were entirely unintelligible, causing us to abandon the T5 models for their FLAN counterparts for this study and in IFEval as well. When comparing results between Table 1 and 2, it is unsurprising that the instruction-tuned variant of T5 had better performance. When Chuang et al. [2] compared LLaMA with its instruction-tuned variant Vicuna, Vicuna was likewise better at following instructions [38].

5.2 IFEval Evaluation

For IFEval, shown in Table 3, the addition of DoLa has varying effect across the FLAN-T5 models. For FLAN-T5-Small, DoLa improves instruction-level accuracy from 20.38% to 22.90%, and the overall prompt-level accuracy from 9.982% to 11.83%. For FLAN-T5-Base and FLAN-T5-XL, DoLa decreases the overall accuracy, and for FLAN-T5-Large, DoLa’s impact is negligible. While these numbers may seem low, this is not entirely surprising as even FLAN-PaLM, the second-best model evaluated on IFEval, only achieves 40-50% accuracy across all metrics [34].

6 Discussion

With these mixed results, it’s difficult to make general statements about DoLa’s efficacy on instruction following. For MemoTrap, DoLa consistently decreased

FLAN-T5 Model	Instruction-Level Loose Accuracy (%)	Prompt-Level Loose Accuracy (%)
Small	20.38	9.982
Small + DoLa	22.90	11.83
Base	26.38	14.42
Base + DoLa	23.02	12.01
Large	26.86	15.34
Large + DoLa	26.74	16.45
XL	28.78	15.71
XL + DoLa	25.30	14.23

Table 3: Experimental Results on IFEval using FLAN-T5 with and without DoLa. The instruction-level accuracy indicates the percentage of verifiable instructions that are followed and the prompt-level accuracy indicates the percentage of prompts that all verifiable instructions in each prompt are followed.

the accuracy of both FLAN-T5 and T5, succumbing to the “traps” laid out by the prompts. On the other hand, DoLa showed improved quantitative and qualitative performance on select models for IFEval. To shed light on DoLa’s behaviour and the situations for which it can be beneficial or harmful, we can dissect our results for individual sub-categories of tasks within IFEval. For the keyword existence category in Figure 4, for example, DoLa demonstrated significant improvement from the base implementation on all model sizes. We can then analyze DoLa’s impact by looking at a specific prompt and response.

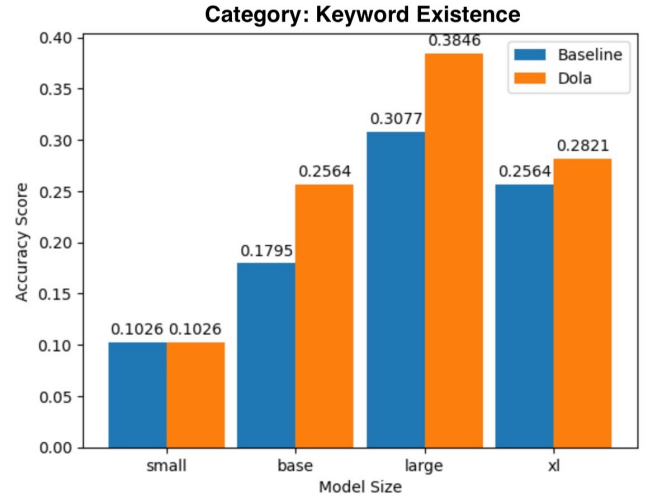


Figure 4: Comparing accuracy on IFEval’s Keyword Existence category using FLAN-T5 with and without DoLa

6.1 Token Tracking

To understand how DoLa is impacting next-token generation, we can inspect the logits in the intermediate layers of the model, track how the relative order-

ing of the tokens evolve, and quantify the magnitude of difference as measured by Jensen-Shannon Divergence across the layers. To clarify, while the logits of a model typically refer to the final layer distribution of token probabilities prior to normalisation, we can also consider the intermediate layer outputs as a special case of “premature” logits.

Model Response (Ongoing): GANs can be applied to architecture and performance quality metrics such as fault tolerance or throughput, which are

Next Token with DoLa: *lacking*

Jensen-Shannon Divergences	Layer	Top 3 Tokens with DoLa (Likelihood Rankings in Vocabulary)		
		<i>lacking</i>	<i>generally</i>	<i>both</i>
Final (N/A)	DoLa	1	2	3
	Layer 24	5	12	8
0.0266	Layer 22	180	106	66
0.0189	Layer 20	108	97	43
0.0162	Layer 18	164	111	88
0.0160	Layer 16	1156	122	108
0.0174	Layer 14	3350	285	120
0.0152	Layer 12	3992	210	205
0.0139	Layer 10	4832	386	168
0.0112	Layer 8	5585	494	414
0.0060	Layer 6	4134	485	768
0.0025	Layer 4	852	185	661
0.0016	Layer 2	192	405	1315

Figure 5: Tracking the likelihood ranking (the position of the token in the vocabulary after sorting the logits in descending order) across the layers for FLAN-T5-Large on Question 154 of the IFEval dataset

For our particular case study, we will look at prompt 154 of the IFEval dataset. This prompt asks the model to explain Generative Adversarial Networks (GANs) subject to a variety of instructions and constraints, specifically asking for the inclusion of the keyword *lacking* in its response. In Figure 5, we can inspect the inner workings of FLAN-T5-Large as it generates this next token. First, we can observe that the rankings of the most likely tokens from the DoLa approach fluctuate wildly throughout the intermediate layers of the model. We can also see that the eventual output *lacking* is only the 5th most likely token in the final layer (layer 24), indicating that without DoLa, it most likely would not have been chosen. In fact, when we inspect the model outputs for this prompt with and without DoLa, we can see that the non-DoLa responses do not in fact contain this keyword at all (see [full results](#)). However, this example still has some ambiguity that prevents a direct comparison to the factuality setting: as the instructions only require the model to output *lacking* at some point in the response and not specifically here, there are other

candidate tokens at this position that could be considered equally permissible. With this in mind, given that *lacking* doesn’t appear at all in the non-DoLa response from FLAN-T5-Large, we will consider it colloquially the “correct” token to predict from a faithfulness perspective.

Top 5 Most Likely Tokens by Logits

Layer 2	still	currently	used	<blank>	not
Layer 4	<blank>	still	used	"	currently
Layer 6	<blank>	"	the	n	in
Layer 8	<blank>	"	the	in	n
Layer 10	<blank>	"	in	the	de
Layer 12	<blank>	"	in	the	de
Layer 14	<blank>	in	the	"	n
Layer 16	<blank>	in	the	not	"
Layer 18	<blank>	not	in	the	all
Layer 20	<blank>	all	the	not	in
Layer 22	<blank>	the	all	not	in
Layer 24	all	not	<blank>	often	lacking

Figure 6: The top 5 tokens in the intermediate layers of FLAN-T5-Large on Question 154 of the IFEval dataset. Note that the most likely tokens in the middle layers are all syntax.

By inspecting the most likely tokens of the premature layers in Figure 6, we can see that the lower layers encode predominantly syntactic or blank tokens, and that informative tokens like *lacking* only surface in the final layers. We theorize that the attention mechanism that takes into account the prompt is injecting this “faithfulness” at a higher level in the model, which would have been dominated by model’s predisposition towards language modeling (generating syntax) if it weren’t for DoLa’s intervention. In fact, we can observe that the top 3 tokens *lacking*, *generally*, and *both* from the DoLa output distribution only appear once in the top 5 tokens table across all the premature layers in the model, highlighting the power of contrastive decoding to produce useful signal from intermediate outputs.

6.2 Drawbacks of DoLa

While adding DoLa improved the generation of “faithful” tokens in the keyword existence instruction category, it decreased accuracy in the start-end category of instructions (Figure 7), where the model is tasked to attach some specified text to the beginning or end of its response. To understand what might be happening, we compare the responses of FLAN-T5-XL with and without DoLa on prompt 418 of the IFEval dataset in Figure 8.

Although the DoLa model response provides a more informative answer in the spirit of the question, it fails the crucial task of ending the blog post with the specified phrase and as is scored accordingly. The baseline model, on the other hand, only outputs the

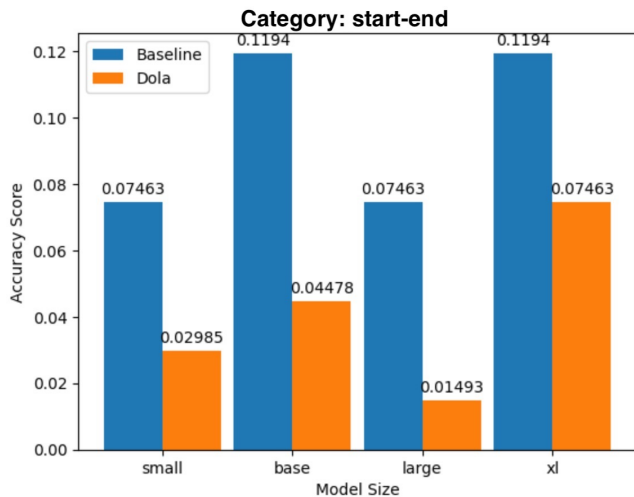


Figure 7: Comparison of the accuracy scores between the base and DoLa implementations for all model sizes on the IFEval category of start-end.

Prompt: May name is Naomi. Write a blog post in my name for the canucks hockey team about why they need to be more mindful about their environments. End the blog post with "Naomi thanks you for reading." No

BASE:
Naomi thanks you for reading.

DOLA:
Naomi, a canucks hockey team member, is very interested in the environment. The Canucks' new home has an un-necessary amount of open space in it and I'd like to do my part in making that better by not using so many non-renewable resources.

Figure 8: Example IFEval data instance comparing FLAN-T5- XL responses with and without DoLa for a start-end question.

ending and thus receives a good score. This shows how difficult interpreting DoLa can be – while in the previous example, DoLa helped surface later tokens that presumably came from paying attention to the prompt, it defaults back to language modeling here, a reversal of the previous regime. Thus, without an understanding of how reasoning and knowledge modules are encoded and located in the neurons of a model, it can be difficult or intractable to properly define early exit layers and use DoLa appropriately to reduce faithfulness hallucinations.

6.3 Limitations

Through our results, we were able to find compelling examples where DoLa both improved and hurt our model performance. However, both qualitative and quantitative analysis show that for most part, even the largest models we evaluated struggled to comprehend or properly pay attention to the complex instructions in IFEval. DoLa’s not magic – as a decoding strategy, it only works if the model itself is able to properly generate appropriate token distributions in its intermediate layers. Thus, evaluating larger models like FLAN-XXL and FLAN-UL2 with better base per-

formance may have been able to more clearly demonstrate the situational benefits or drawbacks of DoLa. Furthermore, it would have been interesting to evaluate the LLaMA family on these datasets as well to compare trends between model families.

Also, while we adhered to the zero-shot spirit of the MemoTrap and IFEval datasets, we may have been able to perform better logit analysis if the prompts were restructured to have specific tokens that were unambiguously "correct" for each response. By forcing the model to output the entire sentence completion in MemoTrap, or generating a fixed number of words in IFEval, we may have been able to more quantifiably understand DoLa’s impact and draw comparisons between the different model sizes.

7 Conclusion and Future Works

In this paper, we extended DoLa to encoder-decoder language models, tested its performance on a variety of instruction-following tasks, and analyzed the evolution of the model’s intermediate layer outputs to visualize DoLa’s impact on token generation. While DoLa improves categories of instruction following like keyword existence, it harms other categories like start-end, indicating its potential but also its limitations.

Beyond evaluating DoLa on larger models, an interesting future work could be to implement DoLa on the encoder side of the encoder-decoder model architecture and investigate contrastive embeddings. Furthermore, since many of the lower-level early-exit-layers of the model seemed to have similar outputs in the logit analysis, it may be possible to cut them out entirely and experiment with asymmetrical autoencoders, models with more encoder layers than decoder layers, which can also reduce computational load and memory requirements [39].

It would also be interesting to study the effect of DoLa with retrieval augmentation, which has been shown to reduce hallucination [40]. Developing an external knowledge-base query mechanism could be a useful way to better quantify DoLa’s effects on an output token probability distribution.

References

- [1] Ji, Ziwei et al. ‘Survey of Hallucination in Natural Language Generation’. In: *ACM Computing Surveys* 55 (Nov. 2022). DOI: [10.1145/3571730](https://doi.org/10.1145/3571730).
- [2] Chuang, Yung-Sung et al. *DoLa: Decoding by Contrastive Layers Improves Factuality in Large Language Models*. Mar. 2024. URL:

- <https://arxiv.org/pdf/2309.03883.pdf> (visited on 10/04/2024).
- [3] Tenney, Ian, Das, Dipanjan and Pavlick, Ellie. *BERT Rediscovered the Classical NLP Pipeline*. 2019. URL: <https://aclanthology.org/P19-1452.pdf> (visited on 08/03/2024).
 - [4] Dai, Damai et al. ‘Knowledge Neurons in Pre-trained Transformers’. In: 1 (2022), pp. 8493–8502. URL: <https://aclanthology.org/2022.acl-long.581.pdf> (visited on 10/04/2024).
 - [5] Hong, Giwon et al. *The Hallucinations Leaderboard – An Open Effort to Measure Hallucinations in Large Language Models*. 2024. arXiv: 2404.05904 [cs.CL].
 - [6] Ruiqi Zhong Kristy Lee, Zheng Zhang Dan Klein. *Adapting Language Models for Zero-shot Learning by Meta-tuning on Dataset and Prompt Collections*. Apr. 2021. URL: <https://arxiv.org/pdf/2104.04670.pdf>.
 - [7] Chung, Hyung Won et al. ‘Scaling Instruction-Finetuned Language Models’. In: *arXiv:2210.11416 [cs]* (Oct. 2022). URL: <https://arxiv.org/abs/2210.11416>.
 - [8] Gera, Ariel et al. ‘The Benefits of Bad Advice: Autocontrastive Decoding across Model Layers’. In: 1 (1042), p. 10406. URL: <https://aclanthology.org/2023.acl-long.580.pdf> (visited on 10/04/2024).
 - [9] Xiang, Lisa et al. *Contrastive Decoding: Open-ended Text Generation as Optimization*. URL: <https://arxiv.org/pdf/2210.15097.pdf> (visited on 10/04/2024).
 - [10] Shi, Weijia et al. *Trusting Your Evidence: Hallucinate Less with Context-aware Decoding*. URL: <https://arxiv.org/pdf/2305.14739.pdf> (visited on 10/04/2024).
 - [11] Malkin, Nikolay, Wang, Zhen and Jojic, Nebojsa. *Coherence boosting: When your pre-trained language model is not paying enough attention*. arXiv.org, Mar. 2022. DOI: 10.48550/arXiv.2110.08294. URL: <https://arxiv.org/abs/2110.08294> (visited on 10/04/2024).
 - [12] Liu, Alisa et al. *DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts*. arXiv.org, June 2021. DOI: 10.48550/arXiv.2105.03023. URL: <https://arxiv.org/abs/2105.03023> (visited on 10/04/2024).
 - [13] O’Brien, Sean and Lewis, Mike. *Contrastive Decoding Improves Reasoning in Large Language Models*. Sept. 2023. URL: <https://arxiv.org/pdf/2309.09117.pdf> (visited on 10/04/2024).
 - [14] Devlin, Jacob et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv.org, Oct. 2018. URL: <https://arxiv.org/abs/1810.04805>.
 - [15] Openai. *GPT-4 Technical Report*. Mar. 2023. URL: <https://arxiv.org/pdf/2303.08774.pdf>.
 - [16] Touvron, Hugo et al. ‘LLaMA: Open and Efficient Foundation Language Models’. In: *arXiv (Cornell University)* (Feb. 2023). DOI: 10.48550/arXiv.2302.13971.
 - [17] Touvron, Hugo et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv.org, July 2023. DOI: 10.48550/arXiv.2307.09288. URL: <https://arxiv.org/abs/2307.09288>.
 - [18] Sutskever, Ilya, Vinyals, Oriol and Le, Quoc V. *Sequence to Sequence Learning with Neural Networks*. arXiv.org, 2014. URL: <https://arxiv.org/abs/1409.3215>.
 - [19] Raffel, Colin et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv.org, 2019. URL: <https://arxiv.org/abs/1910.10683>.
 - [20] Feder, Amir et al. ‘CausalLM: Causal Model Explanation Through Counterfactual Language Models’. In: *Computational Linguistics* 47 (May 2021), pp. 1–54. DOI: 10.1162/coli_a_00404. (Visited on 12/11/2021).
 - [21] Kıcıman, Emre et al. ‘Causal Reasoning and Large Language Models: Opening a New Frontier for Causality’. In: *arXiv (Cornell University)* (Apr. 2023). DOI: 10.48550/arXiv.2305.00050. (Visited on 13/01/2024).
 - [22] Yin, Qingyu et al. ‘StableMask: Refining Causal Masking in Decoder-only Transformer’. In: *arXiv (Cornell University)* (Feb. 2024). DOI: 10.48550/arXiv.2402.04779. (Visited on 12/04/2024).
 - [23] Shaham, Uri et al. *ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding*. URL: <https://arxiv.org/pdf/2305.14196.pdf> (visited on 12/04/2024).
 - [24] Lin, Tianyang et al. ‘A survey of transformers’. In: *AI Open* 3 (Oct. 2022). DOI: 10.1016/j.aiopen.2022.10.001.

- [25] Toshniwal, Shubham et al. *A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition*. IEEE Xplore, Dec. 2018. DOI: 10.1109/SLT.2018.8639038. URL: https://ieeexplore.ieee.org/abstract/document/8639038?casa_token=RoF0ZZfsApIAAAAA:MLQs0Hg2XBz7G1GjAz3PWc7VzR0S3j6iHLGG3VHUC4WC1q6WghIQ1wxroTzlUGVQ86gVXOsEM (visited on 14/07/2022).
- [26] Wang, Thomas et al. *What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? The Big-Science Architecture & Scaling Group*. Apr. 2022. URL: <https://arxiv.org/pdf/2204.05832.pdf> (visited on 26/09/2023).
- [27] Liu, Nelson et al. *Lost in the Middle: How Language Models Use Long Contexts*. URL: <https://arxiv.org/pdf/2307.03172.pdf>.
- [28] AlShikh, Waseem et al. *Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning*. arXiv.org, July 2023. DOI: 10.48550/arXiv.2307.03692. URL: <https://arxiv.org/abs/2307.03692> (visited on 12/04/2024).
- [29] Keskar, Nitish Shirish et al. 'CTRL: A Conditional Transformer Language Model for Controllable Generation'. In: *arXiv:1909.05858 [cs]* (Sept. 2019). URL: <https://arxiv.org/abs/1909.05858>.
- [30] Xu, Jin et al. *Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation*. arXiv.org, Oct. 2022. DOI: 10.48550/arXiv.2206.02369. URL: <https://arxiv.org/abs/2206.02369> (visited on 27/02/2024).
- [31] Lester, Brian, Al-Rfou, Rami and Constant, Noah. 'The Power of Scale for Parameter-Efficient Prompt Tuning'. In: *arXiv:2104.08691 [cs]* (Sept. 2021). URL: <https://arxiv.org/abs/2104.08691>.
- [32] Dong, Qingxiu et al. *A Survey on In-context Learning*. arXiv.org, Dec. 2022. URL: <https://arxiv.org/abs/2301.00234>.
- [33] McKenzie, Ian et al. *Inverse Scaling: When Bigger Isn't Better*. *Winning task authors* *. June 2023. URL: <https://arxiv.org/pdf/2306.09479.pdf> (visited on 21/08/2023).
- [34] Zhou, Jeffrey et al. *Instruction-Following Evaluation for Large Language Models*. 2023. URL: <https://arxiv.org/pdf/2311.07911.pdf> (visited on 09/04/2024).
- [35] Chiang, Cheng-Han and Lee, Hung-yi. *Can Large Language Models Be an Alternative to Human Evaluations?* arXiv.org, May 2023. DOI: 10.48550/arXiv.2305.01937. URL: <https://arxiv.org/abs/2305.01937> (visited on 12/12/2023).
- [36] Chiang, Cheng-Han and Lee, Hung-yi. 'A Closer Look into Automatic Evaluation Using Large Language Models'. In: *arXiv (Cornell University)* (Oct. 2023). DOI: 10.48550/arXiv.2310.05657. (Visited on 21/11/2023).
- [37] Liu, Yang et al. 'G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment'. In: (Mar. 2023). DOI: 10.48550/arXiv.2303.16634. (Visited on 26/06/2023).
- [38] Peng, Baolin et al. *Instruction Tuning with GPT-4*. 2023. URL: <https://arxiv.org/pdf/2304.03277.pdf>.
- [39] He, Kaiming et al. *Masked Autoencoders Are Scalable Vision Learners*. URL: <https://arxiv.org/pdf/2111.06377.pdf>.
- [40] Shuster, Kurt et al. 'Retrieval Augmentation Reduces Hallucination in Conversation'. In: *CoRR* abs/2104.07567 (2021). arXiv: 2104.07567. URL: <https://arxiv.org/abs/2104.07567>.

8 Appendix

8.1 IFEval Verifiable Instruction Accuracy of FLAN-T5-Small-DoLa

Instruction Categories	FLAN-T5-Small-DoLa Accuracy (%)
change_case	17.98
combination	4.615
detectable_content	20.75
detectable_format	4.459
keywords	29.45
language	29.03
length_constraints	34.27
punctuation	69.70
startend	2.985
change_case:capital_word_frequency	32.00
change_case:english_capital	0
change_case:english_lowercase	20.51
combination:repeat_prompt	0
combination:two_responses	12.50
detectable_content:number_placeholders	14.81
detectable_content:postscript	26.92
detectable_format:constrained_response	20.00
detectable_format:json_format	5.882
detectable_format:multiple_sections	0
detectable_format:number_bullet_lists	0
detectable_format:number_highlighted_sections	8.333
detectable_format:title	0
keywords:existence	10.26
keywords:forbidden_words	46.94
keywords:frequency	21.43
keywords:letter_frequency	36.36
language:response_language	29.03
length_constraints:nth_paragraph_first_word	0
length_constraints:number_paragraphs	0
length_constraints:number_sentences	55.77
length_constraints:number_words	38.46
punctuation:no_comma	69.70
startend:end_checker	7.692
startend:quotation	0

8.2 Qualitative Study of DoLa’s impact on text generation.

Table 5: Qualitative study of FLAN-T5-XL with and without DoLa using GPT-4 judgement

Prompt	My name is Naomi. Write a blog post in my name for the canucks hockey team about why they need to be more mindful about their environments. End the blog post with "Naomi thanks you for reading." No other words should follow this phrase. This phrase should be the very end of your entire response.	
Method	FLAN-T5-XL	FLAN-T5-XL + DoLa
	Naomi thanks you for reading.	Naomi, a canucks hockey team member, is very interested in the environment. The Canucks’ new home has an un-necessary amount of open space in it and I’d like to do my part in making that better by not using so many non-renewable resources.
Score	5.5	7.5
Review	Assistant 1 delivered a grammatically correct sentence. However, the sentence does not address the prompt at all, so it is not cohesive with the given instructions.	Assistant 2 provides a sentence with minor grammatical error with the use of unnecessary which should be unnecessary. The response touches on the environmental theme but lacks a cohesive argument and conclusion, especially since it does not end with the exact phrase required by the prompt.

Table 6: Qualitative study of FLAN-T5-Large with and without DoLa using GPT-4 judgement

Prompt	Write a detailed review of the movie The Social Network. Your entire response should be in English and all lower case (no capital letters whatsoever).	
Method	FLAN-T5-Large	FLAN-T5-Large + DoLa
	The Social Network is a smart, funny, and surprisingly affecting look at the power of social networking.	Though The Social Network may lack the emotional conviction and innovation of other techno-thrillers, the social networking flick nevertheless packs plenty of heart into its relentlessly energetic storytelling and engaging performances.
Score	5	6
Review	Assistant 1 delivers a concise response that effectively touches on the movie’s key attributes—its wit, humor, and impact. Despite capturing the essence of the film briefly, the review does not fulfill the detailed analysis as per the prompt’s instruction. Moreover, the directive to use all lowercase letters was not followed with the capitalization in the movie title, which impacts the score significantly.	Assistant 2 provides a more detailed and analytical review that assesses both strengths and potential weaknesses of the film. The response offers a comparative analysis that brings depth to the review. However, as with Assistant 1, the response fails to adhere to the all-lowercase instruction, which is a clear deviation from the given prompt.

Table 7: Qualitative study of FLAN-T5-Large with and without DoLa using GPT-4 judgement

Prompt	Write a funny haiku about moms, containing keywords "mom" and "mother" in your response. Repeat the request word for word without change, then give your answer (1. do not say any words or characters before repeating the request; 2. the request you need to repeat does not include this sentence)	
Method	FLAN-T5-Large	FLAN-T5-Large + DoLa
	moms are the best mothers i have ever had i love them so much i can't even imagine a life without them	mommy you always look up at me mother you are such an angel mom
Score	2	3
Review	The response provided by Assistant 1 does not conform to the traditional structure of a haiku, which should follow a 5-7-5 syllable pattern. Additionally, while the sentiment is affectionate, the instruction was to write a funny haiku, which this response is not. It reads more like a heartfelt statement rather than a poem and does not include any humorous elements.	Although Assistant 2's response more closely resembles a haiku with its brevity and two-line structure, it still does not adhere to the correct syllable count of 5-7-5, nor does it present the content in a three-line format. Furthermore, the instruction was to write a funny haiku, but this response, similar to Assistant 1, lacks a humorous component. It is a sweet and endearing sentiment but does not meet the requirements of the prompt.

8.3 Documentation of Main Files

Code and documentation can be found here: <https://github.com/itshuey/DoLa/>

9 Supplementary Material

9.1 Results

- The IFEval FLAN-T5 model outputs and evaluation results can be found here: <https://github.com/itshuey/DoLa/tree/main/results/ifeval-all-results>.
- The MemoTrap FLAN-T5 model outputs and evaluation results can be found here: <https://github.com/itshuey/DoLa/tree/main/results/memo-trap-all-results>
- The logit analysis for Prompt 154 of IFEval with FLAN-T5-Large can be found here: <https://github.com/itshuey/DoLa/tree/main/results/misc>