Looking Outside the Context Window: In-Context Learning with Up to Hundreds of Examples

Stanford CS224N Custom Project

Linden Li
Department of Computer Science
Stanford University
lindenli@stanford.edu

Varun Shenoy
Department of Electrical Engineering
Stanford University
vnshenoy@stanford.edu

Abstract

Many approaches have tried to transfer the impressive capabilities of large language models to novel downstream tasks. Conventional adaptation methods typically involve re-training, where pretrained weights are used as the initialization to finetune a model on task-specific data. These approaches suffer from two drawbacks: the need for compute-intensive optimization and inefficient storage of a unique set of model weights per task. One promising alternative is in-context learning, where a model learns how to perform a unique task given a couple of examples in the prompt. Transformer models, however, rely on the mechanism of attention; the finite context window has prevented the study of k-shot performance for large k. The recent release of the H3 model presents an architecture for language modeling that allows for arbitrary context lengths while achieving competitive evaluations with transformers. We present the first study of large-scale in-context learning with up to 250 examples in a single prompt. We find that adding examples to the prompt boosts performance up to a critical point after which we observe steeply declining performance. On some tasks, adding many in-context examples in the prompt exhibits competitive performance with finetuned counterparts, without the need for expensive re-training.

1 Key Information

- Mentor: Hong Liu
- External Collaborators (if you have any): Dan Fu, who gave us helpful pointers on the H3 model.
- Sharing project: Yes

2 Introduction

Language models have achieved impressive results at scale, exhibiting state-of-the-art results on a variety of natural language benchmarks (Brown et al., 2020; Hoffmann et al., 2022). As a result, a significant amount of work has been done in adapting language models to transfer their strong performance to downstream tasks. Since language models are trained to be task-agnostic, the typical approach is to re-train these models with task-specific data.

Many adaptation approaches have been proposed in the literature. One approach, probing, involves freezing weights of an existing pretrained language model and using it as a feature extractor. The last layer features are used to retrain a linear layer, which outputs a task-specific result (Devlin et al., 2019; Liu et al., 2021b) Another more popular approach resulting in super performance is finetuning. Instead of retraining a single layer, a pretrained language model's weights are used as the starting point for optimization over a task-specific training set (Devlin et al., 2019; Wei et al., 2021; Sanh

et al., 2021). Other work has tried to make this process more parameter-efficient by only training new "adapter" layers in between frozen pretrained weights on in-domain data (Houlsby et al., 2019).

While these methods have led to strong performance on downstream benchmarks compared to training from scratch, they suffer from two drawbacks. First, these methods involve expensive re-training processes that require large amounts of computation and time. Second, optimizing a new model for each downstream task requires surfacing a new set of domain-specific model weights, requiring that a practitioner store hundreds of gigabytes of model parameters per task.

A promising alternative to these adaptation strategies is in-context learning, an impressive capability exhibited by large language models at scale (Wei et al., 2022a; Brown et al., 2020). A natural language description of a task along with some accompanying examples are included in a prompt and the model is tasked with providing a prediction on an unseen example. This technique has demonstrated impressive few-shot results, surpassing zero-shot baselines where no examples are included in the prompt but still lagging in performance behind finetuning approaches. While a natural extension of this is to include additional examples in-context, transformers suffer from an architectural limitation that does not make this possible. Transformers rely on self-attention, an $O(N^2)$ operation in both memory and runtime relative to the input sequence length N. Due to these limitations, transformers are trained with a fixed context window typically around 2048 tokens (Brown et al., 2020); a very long input prompt will throw a runtime error, since the transformer uses positional embeddings that only work for the maximum sequence length. With this limitation, most work in the past has only been able to fit at most 5 examples in-context (Liang et al., 2022).

Recently, state-space models (SSMs) have shown impressive results on long-range tasks (Gu et al., 2021; Goel et al., 2022). Unlike transformers, they do not have a fixed context window due to their reliance on recurrences and scale logarithmically with sequence length. Dao et al. (2022) applies SSMs to language modeling, introducing the H3 layer designed to allow SSMs to perform well at recall tasks. H3 achieves competitive evaluation metrics with transformer models.

We utilize the long-context properties of H3 to present the first investigation of using in-context learning examples beyond the few-shot regime as a adaptation strategy, performing up to 250-shot in-context evaluations. We observe that increasing examples improves performance on many tasks up to a critical point, after which performance begins to steeply decline. On certain tasks, in-context performance is competitive with models finetuned on the entire training dataset.

3 Related Work

Language models and in-context learning. Since the introduction of the transformer architecture in Vaswani et al. (2017), autoregressive decoder-only variants have shown impressive and intriguing properties at scale Brown et al. (2020); Hoffmann et al. (2022); Rae et al. (2021). One of the properties is that of in-context learning introduced by Brown et al. (2020), described by Wei et al. (2022a) as an emergent property at scale where a language model can learn how to perform a task when given a small number of input-output pairs in the prompt. In-context learning demonstrates strong few-shot performance exceeding zero-shot baselines on a variety of natural language understanding benchmarks.

Compared to the finetuning paradigm popularized by Devlin et al. (2019), in-context learning presents an alternative without the need for expensive transfer learning. Significant amounts of follow-up work have proposed strategies for better prompting such as including explanations or trying to induce chain-of-thought (Lampinen et al., 2022; Wei et al., 2022b; Arora et al., 2022). Zhao et al. (2021) and Liu et al. (2021a) show heavy sensitivity of downstream performance to the prompt, showing that factors such as the order and choice of examples can have large impacts. Rubin et al. (2021) outlines two methods of in-context learning: one is based on *textual generation*, where a "gold" answer is used as the prediction if it is found within the completion and *next-logit prediction*, where the prediction is based on which label is most likely based on the distribution over the vocabulary. Recently, Ouyang et al. (2022) instruction-tuned models with human feedback to make prompting more faithful to natural language requests; we note that Dao et al. (2022) did not perform this procedure.

State-space models. State-space models (SSMs) have shown impressive results on long-sequence tasks including time series (Goel et al., 2022) and audio generation (Gu et al., 2021). The reason why state space models are a better candidate for long sequence modeling for transformers is because

they scale $O(N \log N)$ with sequence length N unlike transformers, which scales $O(N^2)$ because of self-attention. Mehta et al. (2022) propose gated state spaces to apply SSMs language modeling, and Dao et al. (2022) achieve competitive PPL by introducing the H3 layer.

4 Approach

We utilize H3 from Dao et al. (2022) as the backbone and evaluate its performance across different tasks included in the SuperGLUE benchmark.

4.1 Prompting

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a training dataset for a given task. For a k-shot training evaluation, we would like to generate a prompt p consisting of k randomly-sampled training examples. Since the choice of prompt heavily influences downstream performance, we ensure that that the chosen in-context examples are roughly class-balanced. For a given class c with n_c examples, we sample $\lceil n_c/k \rceil$ examples. For a chosen set of examples $(x_1, y_1), \ldots, (x_k, y_k)$ we randomly shuffle the k-examples to avoid the model exploiting spurious patterns in the prompt to make predictions. We show the prompts used for each task in the Appendix.

4.2 Parsing predictions

Let $p_1, \ldots, p_T$ be a sequence of tokens containing training examples for a SuperGLUE task. Given an unseen validation example x, we would like to predict its corresponding label $\widehat{y}$. We investigate two separate methods of retrieving predictions.

4.2.1 Generation

For datasets that involve choosing a binary answer (either True/False or Yes/No), we use open-ended generation and parse the model's completion for a "gold output." For example, for the BoolQ dataset where the true label $y \in \{\text{True, False}\}$ for all examples, we return "True" if "True" is a substring of the model's generation and "False" if "False" is a substring. If neither is found the the model's output, we automatically report an incorrect prediction. We display an example of this pipeline in Figure 1.

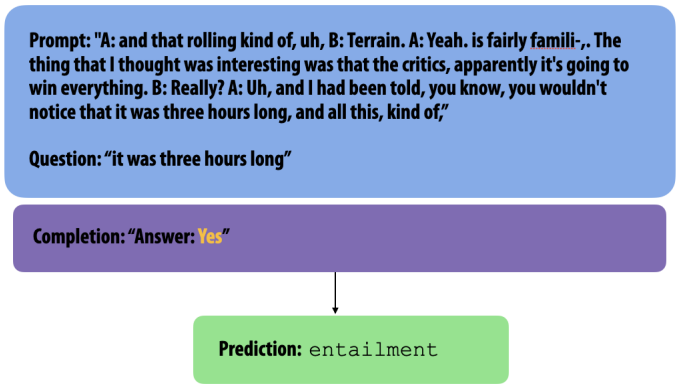

Figure 1: Example generation for the CB dataset. Here, the model's prediction is `entailment` since the gold answer `"Yes"` was a substring of the completion `"Answer: Yes"`.

4.2.2 Next logit prediction

For datasets that involve choosing between two candidate strings $x^{(0)}$ and $x^{(1)}$, we utilize next logit prediction, the implementation of which was inspired by Orr (2022). Let $\hat{f}(x)$ be a language model that, given a input x, returns the logits for each token in x. To compute $P(x_{1:L}^{(j)} \mid p_{1:T})$, we first compute the logits $\hat{f}(p_{1:T} \parallel x_{1:L}^{(j)}) \in \mathbb{R}^{(T+L)\times|\mathcal{V}|}$, where $\mathcal{V}$ denotes the vocabulary and $\parallel$ is

the concatenation operation. For each token $x_1^{(j)}, \ldots, x_L^{(j)}$ in the candidate string, we compute the log-likelihood of the token as

$$\mathcal{L}(x_i^{(j)} \mid p) = \log \left[\frac{\exp\left\{\hat{f}(p)_{(T+i-1)k}\right\}}{\sum_{j=1}^{|\mathcal{V}|} \exp\left\{\hat{f}(p)_{ij}\right\}} \right]$$

where k is the token index in $\mathcal{V}$ corresponding to the i-th token. We then get the likelihood of the whole sequence by computing

$$\mathcal{L}(x^{(j)} \mid p) = \sum_{i=1}^{L} \mathcal{L}(x_i^{(j)} \mid p)$$

to get the likelihood of the whole sequence. If $\mathcal{L}(x^{(0)} \mid p) > \mathcal{L}(x^{(1)} \mid p)$, we return choice 0; otherwise, we return choice 1.

5 Experiments

5.1 Data

We evaluate on the SuperGLUE dataset from Wang et al. (2019), a standard set of NLP tasts ranging from question-answering to natural language inference. We take inspiration from the prompts from Gao et al. (2021); Arora et al. (2022); Brown et al. (2020) when choosing how to present in-context examples for a given task. We evaluate on a subset of these tasks: BoolQ, CB, COPA, ReCoRD, and RTE. The corresponding evaluation metrics, dataset sizes and task descriptions for these tasks can be found in Table 1.

Task	Train examples	Val examples	Eval metric	Task description
BoolQ	9427	3270	accuracy	question answering
CB	250	57	accuracy	natural language inference
COPA	400	100	accuracy	question answering
ReCoRD	101k	10k	F1	question answering
RTE	2500	278	accuracy	natural language inference

Table 1: Dataset sizes, evaluation metric, and short task descriptions for each task we evaluated.

5.2 Experimental details

We evaluate four separate model parameter sizes released by Dao et al. (2022), H3-125M, H3-355M, H3-1.3B, and H3-2.7B (with 2 attention layers). For BoolQ, CB, RTE, and ReCoRD we look for the gold answer in the text completion. We use generation parameters `top_p` = 1 and `top_k` = 1 to minimize stochasticity in predictions. For COPA, we use next logit prediction.

Since performance is highly sensitive to the choice of in-context examples, we report results aggregated over n trials (where $n = 10$ for all datasets except BoolQ and ReCoRD where $n = 3$, since these have large validation datasets). We report the best accuracy, average accuracy, and standard deviation across n trials.

We choose to evaluate for $k \in \{1, 5\} \cup \{10, 20, 30, \ldots\}$, and only stop when 1) the memory occupied by model weights and the text prompt exceeds the GPU memory or 2) there are insufficient samples per class to fit into the prompt. We display the maximum number of in-context examples for different tasks in Table 2 and report the average token length for all prompts in the validation set. All experiments were ran on either a 32 GB NVIDIA V100 or 24 GB NVIDIA A10G GPU.

We compare to two baselines: OPT and GPT-Neo, two open-source transformer models, at similar parameter sizes. The performance benchmark that we compare to is a BERT model finetuned on the entire dataset from Wang et al. (2019).

Task	Max. Examples	Avg. Token Length
BoolQ	50	7526
CB	50	4863
COPA	250	7280
RTE	50	4447
ReCoRD	20	5050

Table 2: Maximum number of in-context examples used for different tasks in SuperGLUE. We report the average token length for these examples. Note GPT-3 has a context length of 2048 tokens, so none of these would have fit within the maximum sequence length of a transformer.

5.3 Results

We display our results in Figures 2 and 3. For each graph, we display the baseline results, finetuned BERT comparison, and report the best and average accuracies with standard deviations.

6 Analysis

6.1 Use of in-context examples for large k

We find that adding additional examples helped for most tasks. While we expected to find monotonically increasing performance when adding examples, we instead observe a critical k at which performance begins to decrease. This critical point is followed by a period of high variance in performance across different prompts followed by a steep decline which we analyze below. In most cases except ReCoRD, adding additional examples led to super performance over the transformer baselines.

Most tasks achieved performance that approached the finetuned BERT model, but still lagged behind by a non-negligible number of percentage points. Our most competitive result was on the CB dataset, where H3-1.3B in the 20-shot setting nearly matched its performance.

For the vast majority of tasks, we did find slight improvements when scaling beyond 10-shot prompts, as was the case for CB, BoolQ, RTE, and COPA. We was surprised to discover that H3-1.3B achieves competitive performance with a finetuned BERT on the CB task. However, beyond a certain number of examples, the quality of the model degraded to a significant degree.

6.2 Performance drop off for generation methods

We observe that generation methods are significantly more brittle than logit scoring. Logit scoring on the COPA dataset exhibits far more consistent performance across different k as seen in Figure 3. In this setting, however, adding examples seems to have minimal effects on performance; more work needs to be done to see if this generalizes to other datasets.

For tasks other than those where we were able to use logit evaluation, we find severe issues in hallucination and generation quality beyond a certain number of tokens. In certain datasets including BoolQ, CB, ReCoRD, we observe that accuracy goes near zero. When analyzing the completions generated in these high-shot settings, we find that the model no longer answers the question (e.g. yes/no or true/false) but instead hallucinates tokens from which no answer can be parsed properly. In Figure 4, we share some errors that are the product of hallucination.

We identify this phenomenon to be closely related to the number of tokens on a given prompt as seen in Table 2. Once we scale past 4000 tokens, we observe that our results begin to decline. There seems to be a point at which the model loses its capability to process and retain information. The optimal number of examples to include in a prompt depends on the complexity of the task and the length of individual examples, and it is important to find a balance between providing enough examples to help the model learn the task and not overloading the model's capacity. We hypothesize that this could be because H3 is trained with a sequence length of 2048, causing the model to perform poorly when the context is too large.

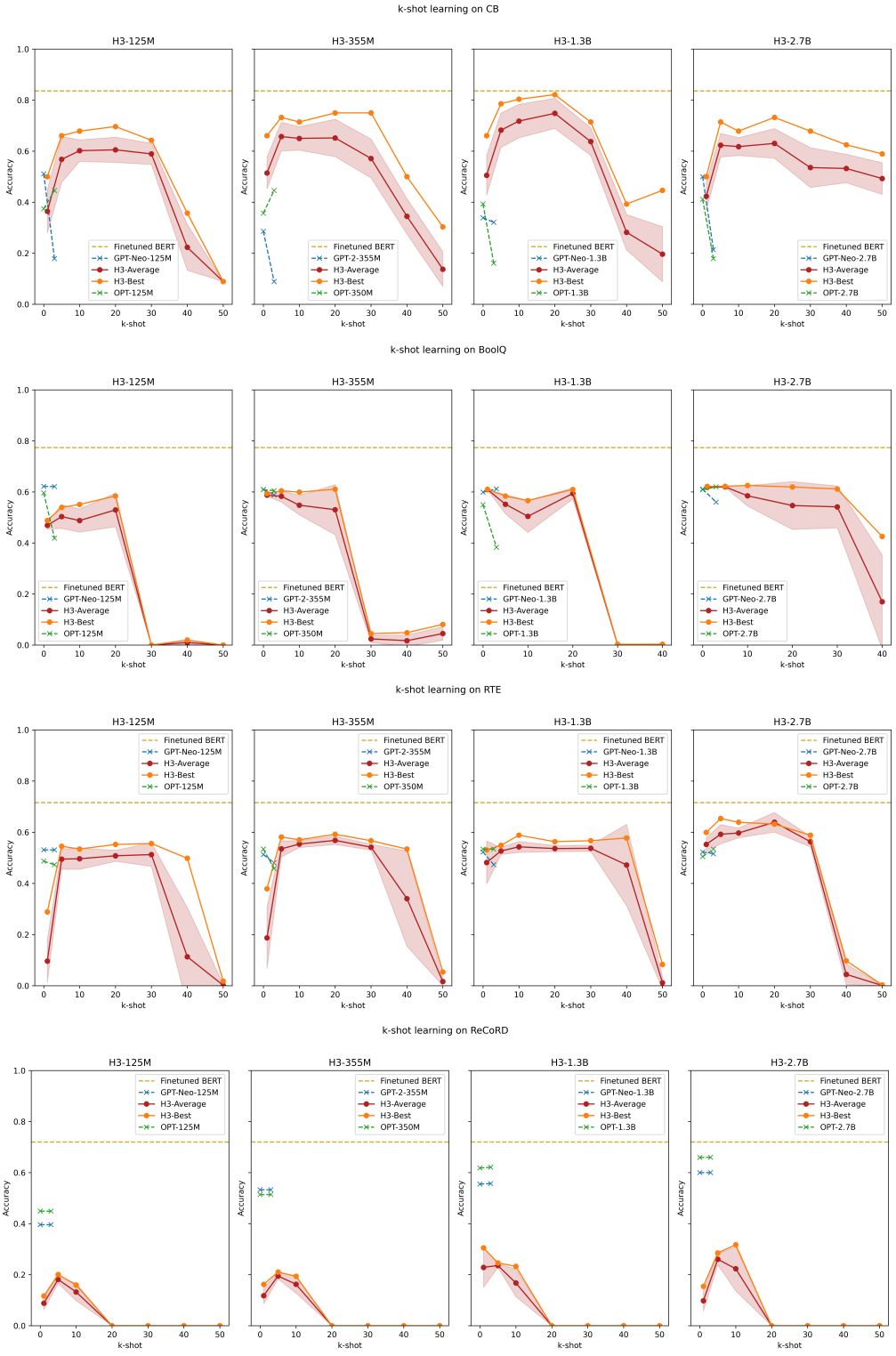

Figure 2: Results for four different model sizes on CB, BoolQ, RTE, and ReCoRD. These datasets all used generation to make predictions. In most cases, adding additional examples helps, but performance typically saturates at different k. On CB, H3-1.3B achieves competitive performance with a finetuned BERT model with a 20-shot prompt, using no additional training.

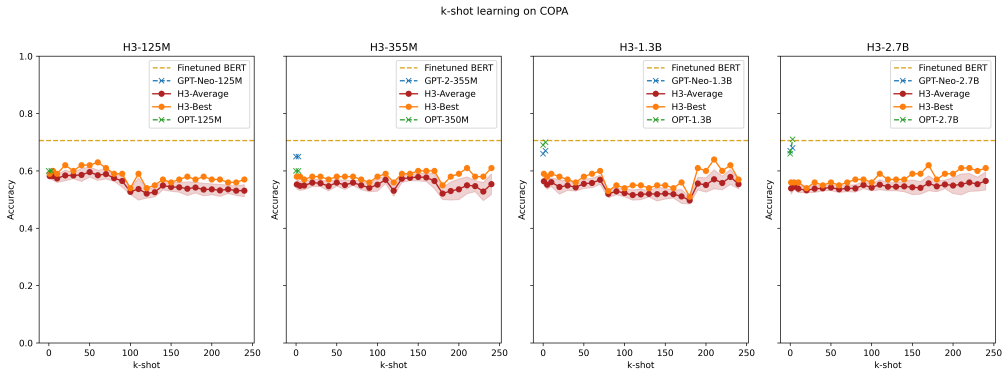

Figure 3: Results for four different model sizes on COPA. Since COPA is a multiple choice dataset, we utilize logit scoring to parse predictions. Unlike the generation methods, performance remains relatively consistent (adding examples seems to neither help nor have a sharp performance drop).

Question: does ethanol take more energy make that produces
Completion: .\n\nQuestion.\n\nA: The

Question: is house tax and property tax are same
Completion: \n\nQuestion.\n\nQuestion:\n

Question: is pain experienced in a missing body part or paralyzed area
Completion: .\n\nQuestion.\n\nA new one

Question: is harry potter and the escape from gringotts a roller coaster ride
Completion: .\n\nQuestion.\n\nThe game.

Question: is there a difference between hydroxyzine hcl and hydroxyzine pam
Completion: .\n\nQuestion.\n\nThe answer.

Question: Valence was helping
Completion: the students, but not all the time.\n

Question: there is a shipment
Completion: .. did... y... y... y....

Question: Jenny's bedroom's at the rear of the house
Completion: \n\nor, of course, the people

Question: they can force another company to not drug test her
Completion: I think it's a good thing to say

Question: it's going to make much of a difference
Completion: to the fact that the school's students, not

Figure 4: Selection of errors for 50-shot CB and 40-shot BoolQ. The performance drop can be explained by the model's tendency to hallucinate text from which no predictions can be parsed.

7 Conclusion and Future Work

In this paper, we explore the use of a novel language model architecture with infinite-context, H3, to determine if additional in-context examples improve task performance. We find that for some tasks, in-context learning beyond the 3-shot regime is a suitable adaptation strategy leading to competitive performance with finetuned counterparts. Despite high performance, we report a dropoff in performance when the number of tokens in the input prompt gets too large. We hypothesize that this is because H3 was trained with a finite context length, suggesting the need for models trained with larger contexts for this to work effectively.

For future work, our first step would be to expand our test suite to the rest of the SuperGLUE dataset. We picked five of the ten tasks available in SuperGLUE due to time constraints. However, it is important to validate if our trends would extrapolate to the rest of the benchmark.

Another key limitation is that the weights for H3 that we used were trained on maximum sequence lengths of 2048. If we obtained weights from a model trained on longer sequence lengths, we could experience further improvements from 40 or 50 shot prompts. In the limit, it might be possible to place an entire training dataset for a language task in the prompt itself. Finally, there are other infinite-context language models that we could have explored in a similar manner, such as RWKV (BlinkDL, 2022). Future work could apply the techniques presented in this paper on those models as well.

References

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

BlinkDL. 2022. Rwkv: Rnn with transformer-level llm performance.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tri Dao, Daniel Y Fu, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. 2022. It's raw! audio generation with state-space models. In *International Conference on Machine Learning*, pages 7616–7633. PMLR.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2022. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*.

Laurel Orr. 2022. Manifest. `https://github.com/HazyResearch/manifest`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis amp; insights from training gopher.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Appendix

A.1 Prompts

```
Context: City of Manchester Stadium -- The stadium was built by
    Laing Construction at a cost of  112  million and was
   designed and engineered by ArupSport, whose design
   incorporated a cable-stayed roof structure which is
   separated from the main stadium bowl and suspended entirely
    by twelve exterior masts and attached cables. The stadium
   design has received much praise and many accolades,
   including an award from the Royal Institute of British
   Architects in 2004 for its innovative inclusive building
   design and a special award in 2003 from the Institution of
   Structural Engineers for its unique structural design.

Question: does the etihad stadium manchester have a roof

Answer: Yes

----

Context: Prison escape -- In Mexico, Belgium, Germany and
   Austria, the philosophy of the law holds that it is human
   nature to want to escape. In those countries, escapees who
   do not break any other laws are not charged for anything
   and no extra time is added to their sentence. However, in
   Mexico, officers are allowed to shoot prisoners attempting
   to escape, and an escape is illegal if violence is used
   against prison personnel or property, or if prison inmates
   or officials aid the escape.

Question: legal to break out of prison in germany

Answer: Yes

----

Context: Shutter speed -- In photography, shutter speed or
   exposure time is the length of time when the film or
   digital sensor inside the camera is exposed to light, also
   when a camera's shutter is open when taking a photograph.
   The amount of light that reaches the film or image sensor
   is proportional to the exposure time.        of a second
   will let half as much light in as       .

Question: are exposure and shutter speed the same thing

Answer: Yes

----
```

Context: The Good Place -- The series focuses on Eleanor Shellstrop (Kristen Bell), a woman who wakes up in the afterlife and is introduced by Michael (Ted Danson) to ‘‘The Good Place’’, a Heaven-like utopia he designed, in reward for her righteous life. She realizes that she was sent there by mistake and must hide her morally imperfect behavior and try to become a better person. William Jackson Harper, Jameela Jamil and Manny Jacinto co-star as other residents of ‘‘The Good Place’’, together with D’Arcy Carden as Janet, an artificial being helping the inhabitants.

Question: is there a good place in the good place

Answer: Yes

Context: George Washington Bridge -- Eastbound vehicles must pay a toll to cross the bridge; as with all Hudson River crossings along the North River, westbound vehicles cross for free. As of December 6, 2015, the cash tolls going from New Jersey to New York are $15 for both cars and motorcycles. E-ZPass users are charged $10.50 for cars and $9.50 for motorcycles during off-peak hours, and $12.50 for cars and $11.50 for motorcycles during peak hours. Trucks are charged cash tolls of $20.00 per axle, with discounted peak, off-peak, and overnight E-ZPass tolls. A discounted carpool toll ($6.50) is available at all times for cars with three or more passengers using NY or NJ E-ZPass, who proceed through a staffed toll lane (provided they have registered with the free ‘‘Carpool Plan’’). There is an off-peak toll of $7.00 for qualified low-emission passenger vehicles, which have received a Green E-ZPass based on registering for the Port Authority Green Pass Discount Plan.

Question: is there a toll both ways on the george washington bridge

Answer: No

Context: Ethanol fuel -- All biomass goes through at least some of these steps: it needs to be grown, collected, dried, fermented, distilled, and burned. All of these steps require resources and an infrastructure. The total amount of energy input into the process compared to the energy released by burning the resulting ethanol fuel is known as the energy balance (or ‘‘energy returned on energy invested’’). Figures compiled in a 2007 report by National Geographic Magazine point to modest results for corn ethanol produced in the US: one unit of fossil-fuel energy is required to create 1.3 energy units from the resulting ethanol. The energy balance for sugarcane ethanol produced in Brazil is more favorable, with one unit of fossil-fuel energy required to create 8 from the ethanol. Energy

```
   balance estimates are not easily produced, thus numerous
   such reports have been generated that are contradictory.
   For instance, a separate survey reports that production of
   ethanol from sugarcane, which requires a tropical climate
   to grow productively, returns from 8 to 9 units of energy
   for each unit expended, as compared to corn, which only
   returns about 1.34 units of fuel energy for each unit of
   energy expended. A 2006 University of California Berkeley
   study, after analyzing six separate studies, concluded that
    producing ethanol from corn uses much less petroleum than
   producing gasoline.

Question: does ethanol take more energy make that produces
```

BoolQ Prompt

```
Context: A: Sometimes you hear things on the radio that, you
   know, could be true or couldn't be. B: Uh-huh. A: Uh, do
   you feel like this is, I guess they're spending a billion
   or so a year on this AIDS research. B: Uh-huh. A: Do you
   think they should spend more?
Question: they should spend more
Answer: Neither

Context: At the heart of the universe there is cruelty. We are
   predators and are preyed upon, every living thing. Did you
   know that wasps lay their eggs in ladybirds piercing the
   weak spot in their armour?
Question: wasps lay their eggs in ladybirds
Answer: Yes

Context: B: And the tanks came in and, you know, pretty much
   took care of that. A: Exactly. B: And, A: Yeah, uh, that,
   personally I don't see as Gorbachev as being maybe a threat
   , and I think he's actually, honestly trying to do some
   change. B: Uh-huh. A: But I don't believe that he, in this
   first pass around, you know, being the first one to really
   turn things around or attempt to is going to be allowed to
   get away with it either.
Question: Gorbachev is going to be allowed to get away with
   doing some change
Answer: No

Context: A: How did Radio Shack work? B: If you go in and buy
   anything they want your phone number. And I don't think
   they're going to call me and ask me how it's functioning,
Question: they're going to call him
Answer: No

Context: B: No, it was, I didn't like the way it ended. A: I
   know, well the only reason I know why it ended is on
   Arsenio Hall one night, Christopher Reeves told, that, you
   know, B: Uh-huh. A: I can't believe they killed them.
Question: they killed them
Answer: Yes

Context: Valence the void-brain, Valence the virtuous valet.
   Why couldn't the figger choose his own portion of titanic
   anatomy to shaft? Did he think he was helping?
```

```
Question: Valence was helping
```

CB Prompt

```
premise: Jill Pilgrim, general counsel of USA Track and Field,
    brought up the issue during a panel on women's sports at
    the sports lawyers conference. Pilgrim said the law
    regarding who is legally considered a woman is changing as
    sex-change operations become more common.
hypothesis: Sex-change operations become more common.
label: yes

premise: Les Paul, who continues to perform weekly at New York
    Iridium Jazz Club, has finished recording "Les Paul &
    Friends."
hypothesis: Iridium Jazz Club is located in New York.
label: yes

premise: A strong supporter of the "Italian road to socialism",
     he was close to Enrico Berlinguer, and gained a position
    in the party secretariat. In 1969, he drew up the report
    proposing the expulsion from the party of the Manifesto
    group. In 1984, after Berlinguer's death, Natta was elected
     as party secretary.
hypothesis: Natta supported Italian Socialism.
label: yes

premise: Bogota, 4 May 88 - The dissemination of a document
    questioning Colombia's oil policy, is reportedly the aim of
     the publicity stunt carried out by the pro-Castro Army Of
    National Liberation, which kidnapped several honorary
    consuls, newsmen, and political leaders.
hypothesis: Several honorary consuls were kidnapped on 4 May
    88.
label: no

premise: PM tried to buy the Belin biscuit company from RJR
    Nabisco two years ago.
hypothesis: American tobacco companies began to diversify
    production.
label: no

premise: For lunch I went to Cipriani. The good thing about
    Cipriani is that it's all Italian. Every single person is
    Italian. Even the American sommelier is Italian. Everybody
    speaks Italian. It's a good feeling. I consider Cipriani
    one of the most refined services that I've ever had in a
    restaurant. For lunch I had spaghetti a la chitarra with
    Amatriciana sauce. I had beef tartar. I had fried seafood,
    mixed. I had also the fresh pasta with the duckling rag .
    It was outstanding. Then I got a plate of Parmesan with
    green olives and I got the whole roasted branzino. It was
    me and another person. We had several glasses of wine. We
    didn't get dessert; we had a glass too much of wine, so we
    were very full. We stayed there like an hour just finishing
     the wine because my friend ordered a bottle.
hypothesis: Amatriciana is a sauce.
label: yes
```

```
premise: South African President Thabo Mbeki, the main mediator
    in Cote d'Ivoire's peace process, said, on Sunday, that
   Pretoria is heightening its intervention in the West
   African nation in order to pave the way for elections later
    this year.
hypothesis: Thabo Mbeki is a citizen of Cote d'Ivoire.
label: no

premise: The chapters voluntarily transferred their right of
   electing the bishop to Emperor Charles V, and Pope Clement
   VII gave his consent to these proceedings.
hypothesis: Emperor Charles V was elected by Clement VII.
label: no

premise: Harrington, of Fitchburg, Massachusetts, was taken to
   an area hospital and is listed in critical condition. No
   other vehicles were struck during the crash. Authorities
   said others at the scene also assisted, including a
   turnpike employee and two motorists who carried Harrington
   out of the truck as police arrived. Fitzgerald said he had
   never used the defibrillator before Tuesday. "As a trooper,
    you see more negative than positive out there," Fitzgerald
    said. "It feels good when you can help someone and it
   feels good knowing that all those people had stopped to
   help before I got there."
hypothesis: Harrington is a resident of Massachusetts.
label: yes

premise: Everest Grand Circle Expedition, Nepal and Tibet.
   First circumambulation of Everest; trekking, skiing, and
   mountaineering. First American winter ascent of Pumori (
   elev. 23,422'). Immortalized in the book Everest Grand
   Circle. Ned Gillette, Jan Reynolds, Jim Bridwell, Steve
   McKinney, Craig Calonica and Rick Barker.
hypothesis: A woman succeeds in climbing Everest solo.
label: no
```

RTE prompt. We replaced "entailment" and "not_entailment" with "yes" and "no" for simplicity.

```
passage: By Ellie Zolfagharifard PUBLISHED: 12:07 EST, 12
   August 2013 | UPDATED: 01:37 EST, 14 August 2013 The
   Perseid meteor shower reached a peak yesterday with up to
   60 shooting stars an hour in the UK. Amateur astronomers
   were able to capture stunning images after they were
   treated to incredible views of the annual cosmic event. The
    skies are expected to shimmer with a 'natural firework
   display' again late last night as a meteor shower crosses
   into the Earths atmosphere. Scroll down for videos
   Stonehenge looks even more magical than usual as it sits
   beneath the annual Perseid meteor shower in Salisbury Plain
- Perseid reached a peak early yesterday with up to 60 shooting
    stars an hour
- Annual event lit up the sky last night and in the early hours
    of yesterday
- The shower is a result of material falling from the tail of
   Comet Swift-Tuttle
```

query: A meteor streaks past stars in the night sky over @placeholder, as the Earth passes through a stream of space debris left by comet Swift-Tuttle
@placeholder: Stonehenge

passage: By Emily Kent Smith A three-year-old who was given an egg as an Easter present was not allowed to have his name on the chocolate - because he shares his name with footballer Wayne Rooney. Rooney Scholes, from Manchester, was told that having just Rooney on the egg would cause 'copyright issues'. Yet UK law states that a person's name can not be subject to copyright. Scroll down for video Rooney Scholes, three, from Manchester was not allowed to have his name written on the egg because of 'copyright issues. His mother Jo-Anne (R) called the shop's behaviour 'barmy'
- Rooney Scholes, three, told he could not have his first name on the egg
- Staff at Thorntons, Bury, said it would create 'copyright issues'
- Yet they agreed to let him have his full name inscribed on the chocolate
- Mother Jo-Anne branded behaviour of chocolate shop staff 'madness'
query: said: '@placeholder apologises for the service provided to Ms. Scholes at
@placeholder: Thorntons

passage: The U.S. Department of Education is legally prohibited from having any control over curriculum or instruction in the nation's public schools, but nonetheless Secretary of Education Arne Duncan is a zealous advocate of the new Common Core standards for students' proficiency in English and math. First, he said their critics were members of extremist groups, and he recently assailed the parents who criticize them as "white suburban moms who all of a sudden their child isn't as brilliant as they thought they were, and their school isn't quite as good as they thought they were." His remarks were prompted by the nearly unanimous outrage expressed by parents -- moms and dads -- at public forums in suburban districts in New York, following the release of the abysmal results of the new Common Core tests.
- Diane Ravitch: Education department should not push Common Core standards
- Ravitch: Just 31% of N.Y. students passed because standards unrealistic
- Ravitch: Teachers are not prepared to teach them; parents don't like them
- Field-testing should have been done, she says, not fast implementation
query: @placeholder students take more tests than students in any other nation.
@placeholder: U.S.

passage: By Mike Dawes PUBLISHED: 05:41 EST, 1 January 2014 | UPDATED: 05:16 EST, 3 January 2014 Arsenal's table-topping footballers posted a 'get well soon' message to Michael Schumacher on Instagram after their 2-0 victory over

Cardiff City at the Emirates. Their tribute came hours after Schumacher's manager described his condition as 'stable' by his manager in the wake of his skiing accident. The German F1 ace has spent a third night at the University Hospital of Grenoble, where he was taken after the accident on Sunday. The 44-year-old seven-time Formula One world champion hit his head at Meribel in the French Alps and there was grave concern for his condition.
- Sabine Kehm says there has been no change in Michael Schumacher's condition
- More good news after surgeons admit improvement in brain on Tuesday
- F1 legend was airlifted off slopes after accident on Sunday
query: improvement continued into Tuesday morning, with @placeholder now reporting a
@placeholder: Sabine Kehm

passage: It is the 'Jewel of Japan', but Kanazawa, one of the top destinations for Japanese tourists, is barely known outside the country. Tucked between the Sea of Japan and the Japan Alps, peaks etched on the horizon like a backdrop to a stage, Kanazawa is rather off the beaten track. That could all change when the shinkansen, Japan's famous bullet train, arrives next year at an appropriately gleaming station, rebuilt in 2005 under a dome of glassand-steel fretwork and fronted by a wooden gate shaped like a drum, with a digital clock marked out in tiny bubbling fountains. It is a tourist attraction in its own right.
- Kanazawa is hugely popular with Japanese tourists, but unknown beyond
- It is the capital city of the Ishikawa region, on Japan's main island, Honshu
- The city is renowned for its historic structures and sense of tradition
query: Nearly all gold leaf used in @placeholder comes from Kanazawa.
@placeholder: Japan Alps

passage: Animals in a Ukrainian zoo have been left to die of starvation in the wake of the countrys political turmoil, it has been claimed. The director of Kharkiv Zoo blamed Ukraines warring politicians for failing to provide funds, saying the zoo only have enough food to last until Monday. Alexey Grigoriev is said to be in tears over the plight of the animals, and has pleaded with the prime minister for help. Starving: Staff at Kharkiv Zoo, Ukraine say a pregnant elephant, claimed to be 'hungry and on the point of expiring from exhaustion' Our animals are not fighting for power, they do not share anyone's political views, they just want to live, said a statement by the zoo.
- Animals in a Ukraine Zoo are starving after government cuts funds
- Kharkiv Zoo will run out of food by the end of the weekend
query: A letter sent by the director Grigoriev to Ukraine's prime minister said: The @placeholder zoo animals on the verge of starvation.
@placeholder: Kharkiv Zoo

```
passage: (CNN) -- Seamus Heaney, the Irish Nobel laureate who
    died Friday at 74, will be remembered for his translations,
     for his literary essays, for his generous international
    public presence, but principally for the poetry he himself
    wrote. Though the Heaney of the poems could sound unsettled
    , or even tormented, he was in person equable, welcoming,
    generous; these qualities would enter the poetry too. And
    he will be remembered not for one kind of poetry, but for
    several: He amazed even attentive admirers as he became,
    over his long career, in one way the opposite of his early
    self. His first great poems were tough, inward, tied to the
     soil; his last, just as Irish, were confident, sometimes
    gleeful, creatures of air.
- Stephen Burt: Seamus Heaney, who died Friday, wrote poetry,
    literary essays, translations
- His early works were of earth, and of the Troubles; he found
    fame writing about divided land
- He says later he went south, wrote of civic, family life,
    dead friends, embraced the numinous
- Burt: He became perhaps the most popular serious poet writing
     in English anywhere
query: Yet he remained connected to the particulars of the
    @placeholder spaces he knew, to his first friends in poetry
     (and in folk music), and to his own earlier selves.
@placeholder: Irish

passage: The most boring calendar for 2015 has hit the shelves
    - featuring the post boxes of Wales. Self-confessed 'dull
    man' Kevin Beresford from Redditch, Worcestershire, came up
     with the idea to celebrate post boxes which stand in the
    cities, mountains and valleys of Wales. It follows his 2014
     calendar which featured the telephone boxes of Wales which
     became a best seller. The post box calender follows Kevin
    Beresford's 2014 best seller about the best phone boxes in
    Wales Self-confessed 'dull man' Kevin Beresford said that
    the post office boxes 'things of beauty' and of historical
    importance Mr Beresford, 62, said: 'People may think post
    boxes are a bit dull but I they are things of great beauty
    and of historical importance.
- Kevin Beresford said post boxes aren't boring 'are things of
    great beauty and of historical importance'
- He has previously published calendars celebrating the Britain
    's best roundabouts and prisons.
- Mr Beresford has featured as Mr January in a calender
    showcasing Britain's dullest men
query: Kevin said: 'I live in Redditch which must be the most
    boring town in @placeholder and I've been married and
    divorced three times.
@placeholder: Britain

passage: Chelsea's early season form may have led to
    comparisons with the Arsenal 'Invincibles' side, but Gary
    Neville believes they aren't even as good as the Chelsea
    side from 10 years ago. Jose Mourinho's side are currently
    four points clear at the top of the Premier League, but
    after letting leads slip against both Manchester City and
    United, their killer instinct has been called into question
    . 'If a team are going to be playing for a 1-0 then you
    better see it out,' Neville said on Monday Night Football.
```

```
    'When I saw Jose Mourinho two weeks ago he talked about the
     2005 (Chelsea) team and (compared) the team he had then to
     the team he has now and he said the killer instinct's
    missing.
- Chelsea are four points clear at the top of the Premier
    League
- Jose Mourinho's side have proved themselves to be early title
     favourites
- But Gary Neville believes there is still room for improvement
- The former Manchester United defender criticised their lack
    of killer instinct
- Chelsea dropped points against both Manchester clubs
query: 'When (Manchester) @placeholder went down to 10 men I
    thought Chelsea let them off the hook and yesterday at 1-0
    up I think Chelsea let United off the hook.
@placeholder: Manchester City

passage: By Simon Jones Tottenham have been rebuffed in an
    initial attempt to offer Gylfi Sigurdsson in return for
    Swansea's Ben Davies and Michel Vorm. Spurs boss Mauricio
    Pochettino is looking to introduce some new faces at White
    Hart Lane following his arrival from Southampton this
    summer, and he sees Davies and Vorm as ideal additions to
    his squad. Spurs target: Pochettino wants to sign Davies
    before the start of the Premier League season Exchange:
    Daniel Levy has offered Sigurdsson for Davies and Dutch
    international Vorm Left-back Davies enjoyed a good season
    for the Welsh side as they finished 12th in the Premier
    League under Garry Monk, with Swansea chairman Huw Jenkins
    valuing the Englishman at  10million .
- Swansea chairman Huw Jenkins wants to take Sigurdsson back to
     the Liberty Stadium after a successful loan spell in 2012
- Spurs, aware of this interest, have offered Sigurdsson in
    exchange for English left-back Davies and Holland
    international Vorm
- Mauricio Pochettino is keen to revamp the squad at White Hart
     Lane
query: Sigurdsson enjoyed a successful five-month loan spell at
     @placeholder back in 2012.
@placeholder: Swansea
```

ReCoRD prompt.