Demonstrations, Not Weights

Task transfer at 220 billion parameters

A report from Brellan AI Research

Summary

A 220 billion parameter autoregressive language model was measured on many tasks with no examples, one example and a handful of examples in its prompt.
With a few demonstrations the model comes close, on many tasks, to systems trained on the task's own data, and no weights change at any point.
Tasks on which it fails are reported beside the ones it passes, together with an analysis of overlap between the training mixture and the benchmarks.

Prepared for software testing. The text is fictional.

Contents

1 Introduction 3
2 Approach 6
3 Results 8
4 Measuring and limiting contamination 9
5 Broader impacts 34
6 Limitations 34
7 Conclusion 36
Appendix A: Details of the models 37
Appendix B: Details of the data 41
Appendix C: Additional results 45

1 Introduction

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.

This report pushes that observation to a much larger scale: a model with 220 billion parameters is measured on many tasks, with no examples, one example, or a handful in the prompt.

The approach is deliberately plain: model and objective stay as they were, and only the prompt's text changes, so that a gain can be credited to the model's use of the examples rather than to a new training method.

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.

1.1 Three settings

Three settings run through the report: zero-shot, with only a description of the task; one-shot, with a single worked example; and few-shot, with a handful of examples, usually between ten and a hundred.

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

Table 1: The eight sizes and their depth and width.

Parameters	Blocks	Width	Heads
110 M	10	640	10
320 M	16	896	12
700 M	20	1280	16
1.4 B	24	1792	24
3.1 B	30	2304	30
7.4 B	36	3584	36
16 B	44	4608	44
220 B	72	10240	72

2 Approach

The pre-training recipe follows our earlier report, with a much larger scale and a broader mixture of text.
The family is autoregressive, training a 220 billion parameter autoregressive language model, which we call Brellan-220B,
and its members are named by parameter count.
Zero-shot scores rise steadily with size, and few-shot scores rise faster,
so the difference between the two settings grows as the model does.
Nothing in these “learning” curves involves a gradient update or a fine-tune — the prompt simply carries more examples.
Brellan-220B, for instance, scores 71.2 F1 on the Quill reading set with no examples, 74.8 with one and 77.5 with several.
Some tasks, however, remain difficult for prompting even at this size.

2.1 Definitions

We use Few-Shot (FS) for the setting in which a handful of worked examples are placed in the prompt and the weights stay frozen
and no task-specific data is used beyond those examples.
One-shot means a single example, and zero-shot means none: only a plain-language description of the task is given.
The number of examples stays small, ten to a hundred, so that the prompt fits the context and the setup resembles what a person could type by hand.

2.2 Training mixture

The training mixture combines web pages, books, articles and code, filtered for breadth and for the quality of the prose.
Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.
Documents are deduplicated, and pages that are mostly boilerplate or that repeat text found elsewhere are dropped.

2.3 Filtering

The web portion was scored by a classifier trained to tell prose resembling a curated corpus from prose that does not, and pages were kept in proportion to that score. The raw crawl is large: 26TB of text before filtering, and 310GB after, which is the version used for training, alongside books, articles and code. Every model in the family saw 210 billion tokens. Our eight sizes span 110 million to 220 billion parameters and share one dataset, so that a single set of runs shows the effect of size. The architecture follows Wrennick-LM, with the changes noted below, which keeps the comparison with the earlier numbers as direct as it can be.

2.4 The training run

Training ran on the Kestrel cluster, which a partner organisation provided,
for the largest models, with smaller ones on the same cluster at proportionally smaller scale.
The mixture was sampled by size rather than by dataset: some parts of the mixture were repeated up to 2.6 times, others were never seen in full.
Benchmarks were checked against the training text. A filtering bug let some of that overlap through, and training the model again at this cost was out of reach,
so the overlap is reported per benchmark, and results are analysed with and without the affected items.

3 Results

Short answers are scored by exact match or accuracy, longer answers by the overlap score F1. On many tasks a handful of examples in the prompt brings the model close to a system trained on the task's labelled set, and sometimes past it.
Confidence is measured as well as accuracy, and the model turns out to be too sure of the answers it gets wrong, more so at larger sizes.

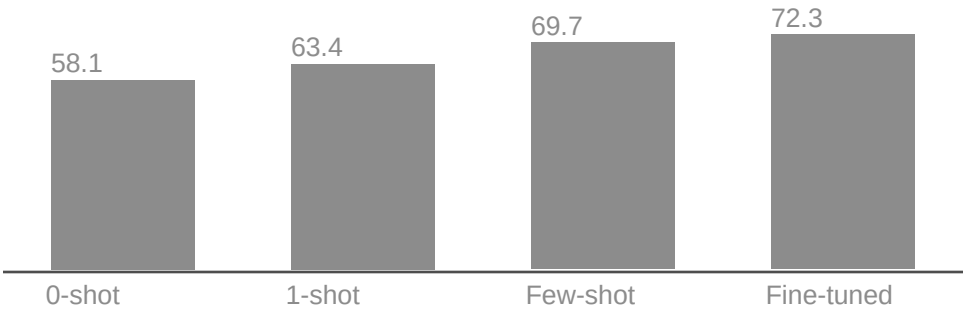
Table 2: Scores on six benchmarks with no examples, one and several, for the largest model.

Benchmark	Zero-shot	One-shot	Few-shot
Reading comprehension	58.1	63.4	69.7
Question answering	49.6	54.2	58.9
Translation	18.3	21.7	23.5
Word puzzles	57.2	66.8	73.1
Common sense	62.4	64.9	67.0
Arithmetic	15.1	18.6	22.9

3.1 Language modelling

Eight sizes are reported, from 110 million to 220 billion parameters, so that each task shows a trend rather than one large-model number.

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.



Score on a reading benchmark by the number of demonstrations in the prompt, beside a fine-tuned system of the previous generation.

3.2 Closed-book question answering

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
This report pushes that observation to a much larger scale: a model with 220 billion parameters is measured on many tasks, with no examples, one example, or a handful in the prompt.
The training mixture combines web pages, books, articles and code, filtered for breadth and for the quality of the prose.

3.3 Reading comprehension

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
Short answers are scored by exact match or accuracy, longer answers by the overlap score F1. On many tasks a handful of examples in the prompt brings the model close to a system trained on the task's labelled set, and sometimes past it.

3.4 Translation

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

Confidence is measured as well as accuracy, and the model turns out to be too sure of the answers it gets wrong, more so at larger sizes.

Eight sizes are reported, from 110 million to 220 billion parameters, so that each task shows a trend rather than one large-model number.

3.5 Winograd schemas

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

A few benchmarks get worse at the largest size. We put that down to prompt format rather than to a loss of ability.

Overlap between the training mixture and each benchmark is measured and reported, since a large overlap would flatter the results.

3.6 Common-sense reasoning

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.
No claim is made that the model reasons about its answers; it sometimes gives the right answer with an explanation a person would not give.
Training compute is reported, and it is large: thousands of accelerators for several weeks.

3.7 Arithmetic

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.
Output can do harm: it can repeat the biases of its training data, invent plausible falsehoods, and produce misleading text at scale.
The measures taken to filter the data and to limit harmful output are described, along with what they cannot do.

3.8 Word scrambling

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.
The model is offered through an interface rather than as weights, so that use can be watched while the risks are studied.
Scale alone buys a great deal of general ability, and returns from a larger model were still positive across the range trained here.

3.9 News generation

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.

Prompt wording is another open question. A small rewording can move a score by several points, which makes cross-paper comparisons awkward.

The report ends with what the results do and do not show, and with evaluations we would run if the data or the metric existed.

3.10 Learning and using tools

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
This report pushes that observation to a much larger scale: a model with 220 billion parameters is measured on many tasks, with no examples, one example, or a handful in the prompt.
The training mixture combines web pages, books, articles and code, filtered for breadth and for the quality of the prose.

4.2 Calibration

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
Short answers are scored by exact match or accuracy, longer answers by the overlap score F1. On many tasks a handful of examples in the prompt brings the model close to a system trained on the task's labelled set, and sometimes past it.

4.3 Prompt sensitivity

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

Confidence is measured as well as accuracy, and the model turns out to be too sure of the answers it gets wrong, more so at larger sizes.

Eight sizes are reported, from 110 million to 220 billion parameters, so that each task shows a trend rather than one large-model number.

4.4 Few-shot versus fine-tuning

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.
A few benchmarks get worse at the largest size. We put that down to prompt format rather than to a loss of ability.
Overlap between the training mixture and each benchmark is measured and reported, since a large overlap would flatter the results.

3.11 Analogical reasoning

On word-relationship puzzles taken from an admissions test, the few-shot score was compared with the scores of people who sat the same test.
The score improved with size, and at the largest size it came close to the human average, which was 48% on the same set.
The comparison is loose: applicants were timed and the model was not.
Prompt wording is another open question. A small rewording can move a score by several points, which makes cross-paper comparisons awkward.

4.1 Measuring overlap

Overlap between the training mixture and each benchmark is measured and reported, since a large overlap would flatter the results.

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.

Table 3: The share of each benchmark whose text appears in training, with scores on the clean and overlapping parts.

Benchmark	Overlap	Clean	Overlapping
Reading comprehension	1.5%	68.9	76.2
Question answering	0.6%	58.2	60.1
Translation	0.0%	23.5	-
Arithmetic	0.0%	22.9	-

5.1 Energy use

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
This report pushes that observation to a much larger scale: a model with 220 billion parameters is measured on many tasks, with no examples, one example, or a handful in the prompt.
The training mixture combines web pages, books, articles and code, filtered for breadth and for the quality of the prose.

5.4 Fairness

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
Short answers are scored by exact match or accuracy, longer answers by the overlap score F1. On many tasks a handful of examples in the prompt brings the model close to a system trained on the task's labelled set, and sometimes past it.

6.2 Broader impacts

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

Confidence is measured as well as accuracy, and the model turns out to be too sure of the answers it gets wrong, more so at larger sizes.

Eight sizes are reported, from 110 million to 220 billion parameters, so that each task shows a trend rather than one large-model number.

6.3 Open questions

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

A few benchmarks get worse at the largest size. We put that down to prompt format rather than to a loss of ability.

Overlap between the training mixture and each benchmark is measured and reported, since a large overlap would flatter the results.

6.4 Evaluation protocol

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

No claim is made that the model reasons about its answers; it sometimes gives the right answer with an explanation a person would not give.

Training compute is reported, and it is large: thousands of accelerators for several weeks.

6.5 Data documentation

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.

Output can do harm: it can repeat the biases of its training data, invent plausible falsehoods, and produce misleading text at scale.

The measures taken to filter the data and to limit harmful output are described, along with what they cannot do.

6.6 Deployment and access

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.

The model is offered through an interface rather than as weights, so that use can be watched while the risks are studied.

Scale alone buys a great deal of general ability, and returns from a larger model were still positive across the range trained here.

6.7 Future work

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.
Prompt wording is another open question. A small rewording can move a score by several points, which makes cross-paper comparisons awkward.
The report ends with what the results do and do not show, and with evaluations we would run if the data or the metric existed.

6.1 Limitations

Text-only training leaves a model with no grounding in video, touch or the physical world, and what it learns comes from writing alone, which leaves out a great deal that a person knows.
Pre-training is also hungry for data: reaching this quality took an enormous corpus.
Learning a new task from a handful of examples is still less reliable than it is for a person.
Output can do harm: it can repeat the biases of its training data, invent plausible falsehoods, and produce misleading text at scale.

5.2 Misuse

Better text synthesis also means more scope for misuse.
The model writes fluent, on-topic and false text at a scale nobody could match, which suits spam, fraud and disinformation.
The measures taken to filter the data and to limit harmful output are described, along with what they cannot do.

5.3 Bias

A model trained on the open web inherits the web's biases, at web scale,
and this one reflects what it read. Gender and occupation are associated across a set of prompts in ways that repeat the stereotypes in the data.
Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.

7 Conclusion

A 220 billion parameter model performs many tasks from a prompt alone, and its quality rises smoothly with parameters and with the number of examples in the prompt.
Failures are reported as fully as successes, and the limits of the overlap analysis are stated.
The results argue for studying scale carefully, and for measuring the risks of large models while they are being built rather than afterwards.

Appendix A: The models

The model is offered through an interface rather than as weights, so that use can be watched while the risks are studied.

Scale alone buys a great deal of general ability, and returns from a larger model were still positive across the range trained here.

Table 4: Training compute and tokens by size.

Parameters	Tokens	Accelerator days
110 M	210 B	0.2
1.4 B	210 B	1.0
16 B	210 B	10.5
220 B	210 B	2,740

Appendix C.1: Additional results

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 5: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
39.1	Benchmark A1	68.0	64.0
39.2	Benchmark B2	81.1	77.3
39.3	Benchmark C3	94.2	90.6
39.4	Benchmark D4	47.3	43.9
39.5	Benchmark E5	60.4	56.2
39.6	Benchmark F6	73.5	69.5
39.7	Benchmark A7	86.6	82.8
39.8	Benchmark B8	39.7	35.1
39.9	Benchmark C9	52.8	48.4

Appendix C.2: Additional results

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 6: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
40.1	Benchmark A1	75.0	71.0
40.2	Benchmark B2	88.1	84.3
40.3	Benchmark C3	41.2	37.6
40.4	Benchmark D4	54.3	50.9
40.5	Benchmark E5	67.4	63.2
40.6	Benchmark F6	80.5	76.5
40.7	Benchmark A7	93.6	89.8
40.8	Benchmark B8	46.7	42.1
40.9	Benchmark C9	59.8	55.4

Appendix C.3: Additional results

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 7: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
41.1	Benchmark A1	82.0	78.0
41.2	Benchmark B2	35.1	31.3
41.3	Benchmark C3	48.2	44.6
41.4	Benchmark D4	61.3	57.9
41.5	Benchmark E5	74.4	70.2
41.6	Benchmark F6	87.5	83.5
41.7	Benchmark A7	40.6	36.8
41.8	Benchmark B8	53.7	49.1
41.9	Benchmark C9	66.8	62.4

Appendix C.4: Additional results

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 8: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
42.1	Benchmark A1	89.0	85.0
42.2	Benchmark B2	42.1	38.3
42.3	Benchmark C3	55.2	51.6
42.4	Benchmark D4	68.3	64.9
42.5	Benchmark E5	81.4	77.2
42.6	Benchmark F6	94.5	90.5
42.7	Benchmark A7	47.6	43.8
42.8	Benchmark B8	60.7	56.1
42.9	Benchmark C9	73.8	69.4

Appendix C.5: Additional results

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 9: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
43.1	Benchmark A1	36.0	32.0
43.2	Benchmark B2	49.1	45.3
43.3	Benchmark C3	62.2	58.6
43.4	Benchmark D4	75.3	71.9
43.5	Benchmark E5	88.4	84.2
43.6	Benchmark F6	41.5	37.5
43.7	Benchmark A7	54.6	50.8
43.8	Benchmark B8	67.7	63.1
43.9	Benchmark C9	80.8	76.4

Appendix C.6: Additional results

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 10: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
44.1	Benchmark A1	43.0	39.0
44.2	Benchmark B2	56.1	52.3
44.3	Benchmark C3	69.2	65.6
44.4	Benchmark D4	82.3	78.9
44.5	Benchmark E5	35.4	31.2
44.6	Benchmark F6	48.5	44.5
44.7	Benchmark A7	61.6	57.8
44.8	Benchmark B8	74.7	70.1
44.9	Benchmark C9	87.8	83.4

Appendix C.7: Additional results

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 11: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
45.1	Benchmark A1	50.0	46.0
45.2	Benchmark B2	63.1	59.3
45.3	Benchmark C3	76.2	72.6
45.4	Benchmark D4	89.3	85.9
45.5	Benchmark E5	42.4	38.2
45.6	Benchmark F6	55.5	51.5
45.7	Benchmark A7	68.6	64.8
45.8	Benchmark B8	81.7	77.1
45.9	Benchmark C9	94.8	90.4

Appendix C.8: Additional results

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 12: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
46.1	Benchmark A1	57.0	53.0
46.2	Benchmark B2	70.1	66.3
46.3	Benchmark C3	83.2	79.6
46.4	Benchmark D4	36.3	32.9
46.5	Benchmark E5	49.4	45.2
46.6	Benchmark F6	62.5	58.5
46.7	Benchmark A7	75.6	71.8
46.8	Benchmark B8	88.7	84.1
46.9	Benchmark C9	41.8	37.4

Appendix C.9: Additional results

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 13: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
47.1	Benchmark A1	64.0	60.0
47.2	Benchmark B2	77.1	73.3
47.3	Benchmark C3	90.2	86.6
47.4	Benchmark D4	43.3	39.9
47.5	Benchmark E5	56.4	52.2
47.6	Benchmark F6	69.5	65.5
47.7	Benchmark A7	82.6	78.8
47.8	Benchmark B8	35.7	31.1
47.9	Benchmark C9	48.8	44.4

Appendix C.10: Additional results

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 14: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
48.1	Benchmark A1	71.0	67.0
48.2	Benchmark B2	84.1	80.3
48.3	Benchmark C3	37.2	33.6
48.4	Benchmark D4	50.3	46.9
48.5	Benchmark E5	63.4	59.2
48.6	Benchmark F6	76.5	72.5
48.7	Benchmark A7	89.6	85.8
48.8	Benchmark B8	42.7	38.1
48.9	Benchmark C9	55.8	51.4

Appendix C.11: Additional results

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 15: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
49.1	Benchmark A1	78.0	74.0
49.2	Benchmark B2	91.1	87.3
49.3	Benchmark C3	44.2	40.6
49.4	Benchmark D4	57.3	53.9
49.5	Benchmark E5	70.4	66.2
49.6	Benchmark F6	83.5	79.5
49.7	Benchmark A7	36.6	32.8
49.8	Benchmark B8	49.7	45.1
49.9	Benchmark C9	62.8	58.4

Appendix C.12: Additional results

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 16: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
50.1	Benchmark A1	85.0	81.0
50.2	Benchmark B2	38.1	34.3
50.3	Benchmark C3	51.2	47.6
50.4	Benchmark D4	64.3	60.9
50.5	Benchmark E5	77.4	73.2
50.6	Benchmark F6	90.5	86.5
50.7	Benchmark A7	43.6	39.8
50.8	Benchmark B8	56.7	52.1
50.9	Benchmark C9	69.8	65.4

Appendix C.13: Additional results

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 17: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
51.1	Benchmark A1	92.0	88.0
51.2	Benchmark B2	45.1	41.3
51.3	Benchmark C3	58.2	54.6
51.4	Benchmark D4	71.3	67.9
51.5	Benchmark E5	84.4	80.2
51.6	Benchmark F6	37.5	33.5
51.7	Benchmark A7	50.6	46.8
51.8	Benchmark B8	63.7	59.1
51.9	Benchmark C9	76.8	72.4

Appendix C.14: Additional results

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 18: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
52.1	Benchmark A1	39.0	35.0
52.2	Benchmark B2	52.1	48.3
52.3	Benchmark C3	65.2	61.6
52.4	Benchmark D4	78.3	74.9
52.5	Benchmark E5	91.4	87.2
52.6	Benchmark F6	44.5	40.5
52.7	Benchmark A7	57.6	53.8
52.8	Benchmark B8	70.7	66.1
52.9	Benchmark C9	83.8	79.4

Appendix C.15: Additional results

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 19: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
53.1	Benchmark A1	46.0	42.0
53.2	Benchmark B2	59.1	55.3
53.3	Benchmark C3	72.2	68.6
53.4	Benchmark D4	85.3	81.9
53.5	Benchmark E5	38.4	34.2
53.6	Benchmark F6	51.5	47.5
53.7	Benchmark A7	64.6	60.8
53.8	Benchmark B8	77.7	73.1
53.9	Benchmark C9	90.8	86.4

Appendix C.16: Additional results

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 20: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
54.1	Benchmark A1	53.0	49.0
54.2	Benchmark B2	66.1	62.3
54.3	Benchmark C3	79.2	75.6
54.4	Benchmark D4	92.3	88.9
54.5	Benchmark E5	45.4	41.2
54.6	Benchmark F6	58.5	54.5
54.7	Benchmark A7	71.6	67.8
54.8	Benchmark B8	84.7	80.1
54.9	Benchmark C9	37.8	33.4

Appendix C.17: Additional results

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 21: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
55.1	Benchmark A1	60.0	56.0
55.2	Benchmark B2	73.1	69.3
55.3	Benchmark C3	86.2	82.6
55.4	Benchmark D4	39.3	35.9
55.5	Benchmark E5	52.4	48.2
55.6	Benchmark F6	65.5	61.5
55.7	Benchmark A7	78.6	74.8
55.8	Benchmark B8	91.7	87.1
55.9	Benchmark C9	44.8	40.4

Appendix C.18: Additional results

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 22: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
56.1	Benchmark A1	67.0	63.0
56.2	Benchmark B2	80.1	76.3
56.3	Benchmark C3	93.2	89.6
56.4	Benchmark D4	46.3	42.9
56.5	Benchmark E5	59.4	55.2
56.6	Benchmark F6	72.5	68.5
56.7	Benchmark A7	85.6	81.8
56.8	Benchmark B8	38.7	34.1
56.9	Benchmark C9	51.8	47.4

Appendix C.19: Additional results

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 23: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
57.1	Benchmark A1	74.0	70.0
57.2	Benchmark B2	87.1	83.3
57.3	Benchmark C3	40.2	36.6
57.4	Benchmark D4	53.3	49.9
57.5	Benchmark E5	66.4	62.2
57.6	Benchmark F6	79.5	75.5
57.7	Benchmark A7	92.6	88.8
57.8	Benchmark B8	45.7	41.1
57.9	Benchmark C9	58.8	54.4

Appendix C.20: Additional results

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 24: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
58.1	Benchmark A1	81.0	77.0
58.2	Benchmark B2	94.1	90.3
58.3	Benchmark C3	47.2	43.6
58.4	Benchmark D4	60.3	56.9
58.5	Benchmark E5	73.4	69.2
58.6	Benchmark F6	86.5	82.5
58.7	Benchmark A7	39.6	35.8
58.8	Benchmark B8	52.7	48.1
58.9	Benchmark C9	65.8	61.4

Appendix C.21: Additional results

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 25: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
59.1	Benchmark A1	88.0	84.0
59.2	Benchmark B2	41.1	37.3
59.3	Benchmark C3	54.2	50.6
59.4	Benchmark D4	67.3	63.9
59.5	Benchmark E5	80.4	76.2
59.6	Benchmark F6	93.5	89.5
59.7	Benchmark A7	46.6	42.8
59.8	Benchmark B8	59.7	55.1
59.9	Benchmark C9	72.8	68.4

Appendix C.22: Additional results

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 26: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
60.1	Benchmark A1	35.0	31.0
60.2	Benchmark B2	48.1	44.3
60.3	Benchmark C3	61.2	57.6
60.4	Benchmark D4	74.3	70.9
60.5	Benchmark E5	87.4	83.2
60.6	Benchmark F6	40.5	36.5
60.7	Benchmark A7	53.6	49.8
60.8	Benchmark B8	66.7	62.1
60.9	Benchmark C9	79.8	75.4

Appendix C.23: Additional results

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 27: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
61.1	Benchmark A1	42.0	38.0
61.2	Benchmark B2	55.1	51.3
61.3	Benchmark C3	68.2	64.6
61.4	Benchmark D4	81.3	77.9
61.5	Benchmark E5	94.4	90.2
61.6	Benchmark F6	47.5	43.5
61.7	Benchmark A7	60.6	56.8
61.8	Benchmark B8	73.7	69.1
61.9	Benchmark C9	86.8	82.4

Appendix C.24: Additional results

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 28: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
62.1	Benchmark A1	49.0	45.0
62.2	Benchmark B2	62.1	58.3
62.3	Benchmark C3	75.2	71.6
62.4	Benchmark D4	88.3	84.9
62.5	Benchmark E5	41.4	37.2
62.6	Benchmark F6	54.5	50.5
62.7	Benchmark A7	67.6	63.8
62.8	Benchmark B8	80.7	76.1
62.9	Benchmark C9	93.8	89.4

Appendix C.25: Additional results

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 29: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
63.1	Benchmark A1	56.0	52.0
63.2	Benchmark B2	69.1	65.3
63.3	Benchmark C3	82.2	78.6
63.4	Benchmark D4	35.3	31.9
63.5	Benchmark E5	48.4	44.2
63.6	Benchmark F6	61.5	57.5
63.7	Benchmark A7	74.6	70.8
63.8	Benchmark B8	87.7	83.1
63.9	Benchmark C9	40.8	36.4

Appendix C.26: Additional results

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 30: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
64.1	Benchmark A1	63.0	59.0
64.2	Benchmark B2	76.1	72.3
64.3	Benchmark C3	89.2	85.6
64.4	Benchmark D4	42.3	38.9
64.5	Benchmark E5	55.4	51.2
64.6	Benchmark F6	68.5	64.5
64.7	Benchmark A7	81.6	77.8
64.8	Benchmark B8	94.7	90.1
64.9	Benchmark C9	47.8	43.4

Appendix C.27: Additional results

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 31: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
65.1	Benchmark A1	70.0	66.0
65.2	Benchmark B2	83.1	79.3
65.3	Benchmark C3	36.2	32.6
65.4	Benchmark D4	49.3	45.9
65.5	Benchmark E5	62.4	58.2
65.6	Benchmark F6	75.5	71.5
65.7	Benchmark A7	88.6	84.8
65.8	Benchmark B8	41.7	37.1
65.9	Benchmark C9	54.8	50.4

Appendix C.28: Additional results

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 32: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
66.1	Benchmark A1	77.0	73.0
66.2	Benchmark B2	90.1	86.3
66.3	Benchmark C3	43.2	39.6
66.4	Benchmark D4	56.3	52.9
66.5	Benchmark E5	69.4	65.2
66.6	Benchmark F6	82.5	78.5
66.7	Benchmark A7	35.6	31.8
66.8	Benchmark B8	48.7	44.1
66.9	Benchmark C9	61.8	57.4

Appendix C.29: Additional results

The same prompt format covers writing code from a description, answering questions about a passage, and translating between languages.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 33: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
67.1	Benchmark A1	84.0	80.0
67.2	Benchmark B2	37.1	33.3
67.3	Benchmark C3	50.2	46.6
67.4	Benchmark D4	63.3	59.9
67.5	Benchmark E5	76.4	72.2
67.6	Benchmark F6	89.5	85.5
67.7	Benchmark A7	42.6	38.8
67.8	Benchmark B8	55.7	51.1
67.9	Benchmark C9	68.8	64.4

Appendix C.30: Additional results

The environmental cost of a run this size is discussed, with the energy used and the offsets bought.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 34: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
68.1	Benchmark A1	91.0	87.0
68.2	Benchmark B2	44.1	40.3
68.3	Benchmark C3	57.2	53.6
68.4	Benchmark D4	70.3	66.9
68.5	Benchmark E5	83.4	79.2
68.6	Benchmark F6	36.5	32.5
68.7	Benchmark A7	49.6	45.8
68.8	Benchmark B8	62.7	58.1
68.9	Benchmark C9	75.8	71.4

Appendix C.31: Additional results

Prompts, answer formats and scoring rules for every benchmark are listed in the appendix so that the results can be repeated.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 35: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
69.1	Benchmark A1	38.0	34.0
69.2	Benchmark B2	51.1	47.3
69.3	Benchmark C3	64.2	60.6
69.4	Benchmark D4	77.3	73.9
69.5	Benchmark E5	90.4	86.2
69.6	Benchmark F6	43.5	39.5
69.7	Benchmark A7	56.6	52.8
69.8	Benchmark B8	69.7	65.1
69.9	Benchmark C9	82.8	78.4

Appendix C.32: Additional results

How much of the gain belongs to the data rather than the parameters is open; an ablation trains a small model on more text and a large one on less, holding compute fixed.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 36: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
70.1	Benchmark A1	45.0	41.0
70.2	Benchmark B2	58.1	54.3
70.3	Benchmark C3	71.2	67.6
70.4	Benchmark D4	84.3	80.9
70.5	Benchmark E5	37.4	33.2
70.6	Benchmark F6	50.5	46.5
70.7	Benchmark A7	63.6	59.8
70.8	Benchmark B8	76.7	72.1
70.9	Benchmark C9	89.8	85.4

Appendix C.33: Additional results

It has become clear that a model trained on a large corpus can be pointed at a new task with a short prompt instead of a change to its weights.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 37: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
71.1	Benchmark A1	52.0	48.0
71.2	Benchmark B2	65.1	61.3
71.3	Benchmark C3	78.2	74.6
71.4	Benchmark D4	91.3	87.9
71.5	Benchmark E5	44.4	40.2
71.6	Benchmark F6	57.5	53.5
71.7	Benchmark A7	70.6	66.8
71.8	Benchmark B8	83.7	79.1
71.9	Benchmark C9	36.8	32.4

Appendix C.34: Additional results

The architecture is autoregressive and the objective is next-token prediction, as in earlier work; there is no task-specific head and no fine-tuning step.
The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 38: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
72.1	Benchmark A1	59.0	55.0
72.2	Benchmark B2	72.1	68.3
72.3	Benchmark C3	85.2	81.6
72.4	Benchmark D4	38.3	34.9
72.5	Benchmark E5	51.4	47.2
72.6	Benchmark F6	64.5	60.5
72.7	Benchmark A7	77.6	73.8
72.8	Benchmark B8	90.7	86.1
72.9	Benchmark C9	43.8	39.4

Appendix C.35: Additional results

The benefit of examples is largest where the answer's shape is easy to copy from them, and smallest where several steps of reasoning are needed.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 39: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
73.1	Benchmark A1	66.0	62.0
73.2	Benchmark B2	79.1	75.3
73.3	Benchmark C3	92.2	88.6
73.4	Benchmark D4	45.3	41.9
73.5	Benchmark E5	58.4	54.2
73.6	Benchmark F6	71.5	67.5
73.7	Benchmark A7	84.6	80.8
73.8	Benchmark B8	37.7	33.1
73.9	Benchmark C9	50.8	46.4

Appendix C.36: Additional results

Bigger models need fewer examples to reach a given score, and the score rises smoothly as examples are added.

The table carries the full results for this group of benchmarks, with no examples and with several, for every size trained.

Table 40: Results for the benchmarks in this group.

Task	Benchmark	Zero-shot	Few-shot
74.1	Benchmark A1	73.0	69.0
74.2	Benchmark B2	86.1	82.3
74.3	Benchmark C3	39.2	35.6
74.4	Benchmark D4	52.3	48.9
74.5	Benchmark E5	65.4	61.2
74.6	Benchmark F6	78.5	74.5
74.7	Benchmark A7	91.6	87.8
74.8	Benchmark B8	44.7	40.1
74.9	Benchmark C9	57.8	53.4

References

[1] A. Author, B. Author and C. Author. Weighting keys for sequence models. A technical report, 2017.

[2] A. Author, B. Author and C. Author. What a corpus teaches. A technical report, 2019.

[3] A. Author and B. Author. Reading comprehension with pretrained models. A technical report, 2019.

[4] A. Author, B. Author and C. Author. Overlap between training and evaluation data. A technical report, 2020.

[5] A. Author. Energy and policy considerations for deep learning. A technical report, 2019.