# GPT Performance

# GPT & Benchmarks

**Learning goals**

- Recap GPT and the ideas behind standard language modelling
- Understand the difference between fine-tuning and X-shot learning

# LAMBADA TASK

| | |
|---|---|
| Context → | Fill in blank:<br><br>She held the torch in front of her.<br><br>She caught her breath.<br><br>"Chris? There's a step."<br><br>"What?"<br><br>"A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ____. -> |
| Target Completion → | step |

# PERFORMANCE ON LAMBADA

| Setting | LAMBADA (acc) | LAMBADA (ppl) | StoryCloze (acc) | HellaSwag (acc) |
|---|---|---|---|---|
| SOTA | 68.0[a] | 8.63[b] | **91.8**[c] | **85.6**[d] |
| GPT-3 Zero-Shot | **76.2** | **3.00** | 83.2 | 78.9 |
| GPT-3 One-Shot | **72.5** | **3.35** | 84.7 | 78.1 |
| GPT-3 Few-Shot | **86.4** | **1.92** | 87.7 | 79.3 |

# “CLOSED BOOK” QUESTION ANSWERING (QA) TASK

| | |
|---|---|
| Context → | Q: ‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?<br><br>A: |
| Target Completion → | MARCEL DUCHAMP |
| Target Completion → | r mutt |
| Target Completion → | duchamp |
| Target Completion → | marcel duchamp |
| Target Completion → | R.Mutt |
| Target Completion → | Marcel duChamp |
| Target Completion → | Henri-Robert-Marcel Duchamp |
| Target Completion → | Marcel du Champ |
| Target Completion → | henri robert marcel duchamp |
| Target Completion → | Duchampian |
| Target Completion → | Duchamp |
| Target Completion → | duchampian |
| Target Completion → | marcel du champ |
| Target Completion → | Marcel Duchamp |
| Target Completion → | MARCEL DUCHAMP |

# PERFORMANCE ON CLOSED-BOOK QA TASK

| Setting | NaturalQS | WebQS | TriviaQA |
|---|---|---|---|
| RAG (Fine-tuned, Open-Domain) [LPP+20] | **44.5** | **45.5** | **68.0** |
| T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20] | 36.6 | 44.7 | 60.5 |
| T5-11B (Fine-tuned, Closed-Book) | 34.5 | 37.4 | 50.1 |
| GPT-3 Zero-Shot | 14.6 | 14.4 | 64.3 |
| GPT-3 One-Shot | 23.0 | 25.3 | **68.0** |
| GPT-3 Few-Shot | 29.9 | 41.5 | **71.2** |

# PERFORMANCE ON MACHINE TRANSLATION

| Setting | En→Fr | Fr→En | En→De | De→En | En→Ro | Ro→En |
|---|---|---|---|---|---|---|
| SOTA (Supervised) | **45.6**[a] | 35.0 [b] | **41.2**[c] | 40.2[d] | **38.5**[e] | **39.9**[e] |
| XLM [LC19] | 33.4 | 33.3 | 26.4 | 34.3 | 33.3 | 31.8 |
| MASS [STQ+19] | 37.5 | 34.9 | 28.3 | 35.2 | 35.2 | 33.1 |
| mBART [LGG+20] | - | - | 29.8 | 34.0 | 35.0 | 30.5 |
| GPT-3 Zero-Shot | 25.2 | 21.2 | 24.6 | 27.2 | 14.1 | 19.9 |
| GPT-3 One-Shot | 28.3 | 33.7 | 26.2 | 30.4 | 20.6 | 38.6 |
| GPT-3 Few-Shot | 32.6 | 39.2 | 29.7 | 40.6 | 21.0 | 39.5 |

# WINOGRAD TASK

| | |
|---|---|
| Correct Context → | Grace was happy to trade me her sweater for my jacket. She thinks the sweater |
| Incorrect Context → | Grace was happy to trade me her sweater for my jacket. She thinks the jacket |
| Target Completion → | looks dowdy on her. |

# PERFORMANCE ON WINOGRAD TASK

| Setting | Winograd | Winogrande (XL) |
|---|---|---|
| Fine-tuned SOTA | **90.1**[a] | **84.6**[b] |
| GPT-3 Zero-Shot | 88.3* | 70.2 |
| GPT-3 One-Shot | 89.7* | 73.2 |
| GPT-3 Few-Shot | 88.6* | 77.7 |

# ARC TASK

| | |
|---|---|
| Context → | Question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?<br>Answer: |
| Correct Answer → | dry palms |
| Incorrect Answer → | wet palms |
| Incorrect Answer → | palms covered with oil |
| Incorrect Answer → | palms covered with lotion |

# PERFORMANCE ON ARC TASK

| Setting | PIQA | ARC (Easy) | ARC (Challenge) | OpenBookQA |
|---|---|---|---|---|
| Fine-tuned SOTA | 79.4 | **92.0**[KKS+20] | **78.5**[KKS+20] | **87.2**[KKS+20] |
| GPT-3 Zero-Shot | **80.5*** | 68.8 | 51.4 | 57.6 |
| GPT-3 One-Shot | **80.5*** | 71.2 | 53.2 | 58.8 |
| GPT-3 Few-Shot | **82.8*** | 70.1 | 51.5 | 65.4 |

# RACE TASK

# RACE TASK

| | |
|---|---|
| Context → | Article:<br>Informal conversation is an important part of any business relationship.Before you start a discussion,however,make sure you understand which topics are suitable and which are considered taboo in a particular culture. Latin Americans enjoy sharing information about their local history, art and customs.You may expect questions about your family,and be sure to show pictures of your children.You may feel free to ask similar questions of your Latin American friends.The French think of conversation as an art form,and they enjoy the value of lively discussions as well as disagreements. For them,arguments can be interesting and they can cover pretty much or any topic ---- as long as they occur in are respectful and intelligent manner.<br>In the United States,business people like to discuss a wide range of topics,including opinions about work,family,hobbies,and politics. In Japan,China,and Korea,however,people are much more private.They do not share much about their thoughts,feelings,or emotions because they feel that doing so might take away from the harmonious business relationship they're trying to build.Middle Easterners are also private about their personal lives and family matters.It is considered rude,for example,to ask a businessman from Saudi Arabia about his wife or children.<br>As a general rule,it's best not to talk about politics or religion with your business friends.This can get you into trouble,even in the United States,where people hold different religious views.In addition,discussing one's salary is usually considered unsuitable.Sports is typically a friendly subject in most parts of the world,although be careful not to criticize national sport.Instead,be friendly and praise your host's team.<br><br>Q: What shouldn't you do when talking about sports with colleagues from another country?<br><br>A: Criticizing the sports of your colleagues' country.<br><br>Q: Which is typically a friendly topic in most places according to the author?<br><br>A: Sports.<br><br>Q: Why are people from Asia more private in their conversation with others?<br><br>A: They don't want to have their good relationship with others harmed by informal conversation.<br><br>Q: The author considers politics and religion _ .<br><br>A: |
| Correct Answer → | taboo |
| Incorrect Answer → | cheerful topics |
| Incorrect Answer → | rude topics |
| Incorrect Answer → | topics that can never be talked about |

# PERFORMANCE ON RACE TASK

| Setting | CoQA | DROP | QuAC | SQuADv2 | RACE-h | RACE-m |
|---|---|---|---|---|---|---|
| Fine-tuned SOTA | **90.7**[a] | **89.1**[b] | **74.4**[c] | **93.0**[d] | **90.0**[e] | **93.1**[e] |
| GPT-3 Zero-Shot | 81.5 | 23.6 | 41.5 | 59.5 | 45.5 | 58.4 |
| GPT-3 One-Shot | 84.0 | 34.3 | 43.3 | 65.4 | 45.9 | 57.4 |
| GPT-3 Few-Shot | 85.0 | 36.5 | 44.3 | 69.8 | 46.8 | 58.1 |

# PERFORMANCE ON SUPERGLUE TASK

| | SuperGLUE Average | BoolQ Accuracy | CB Accuracy | CB F1 | COPA Accuracy | RTE Accuracy |
|---|---|---|---|---|---|---|
| Fine-tuned SOTA | **89.0** | **91.0** | **96.9** | **93.9** | **94.8** | **92.5** |
| Fine-tuned BERT-Large | 69.0 | 77.4 | 83.6 | 75.7 | 70.6 | 71.7 |
| GPT-3 Few-Shot | 71.8 | 76.4 | 75.6 | 52.0 | 92.0 | 69.0 |

| | WiC Accuracy | WSC Accuracy | MultiRC Accuracy | MultiRC F1a | ReCoRD Accuracy | ReCoRD F1 |
|---|---|---|---|---|---|---|
| Fine-tuned SOTA | **76.1** | **93.8** | **62.3** | **88.2** | **92.5** | **93.3** |
| Fine-tuned BERT-Large | 69.6 | 64.6 | 24.1 | 70.0 | 71.3 | 72.0 |
| GPT-3 Few-Shot | 49.4 | 80.1 | 30.5 | 75.4 | 90.2 | 91.1 |

# SUPERGLUE

- BoolQ
- CB (true/false/neither)
- COPA
- RTE (similar to natural language inference)
- WiC
- WSC
- MultiRC (true/false)
- ReCoRD

# BOOLQ (BOOLEAN QUESTION) TASK

| | |
|---|---|
| Context → | Normal force -- In a simple case such as an object resting upon a table, the normal force on the object is equal but in opposite direction to the gravitational force applied on the object (or the weight of the object), that is, N = m g (\displaystyle N=mg), where m is mass, and g is the gravitational field strength (about 9.81 m/s on Earth). The normal force here represents the force applied by the table against the object that prevents it from sinking through the table and requires that the table is sturdy enough to deliver this normal force without breaking. However, it is easy to assume that the normal force and weight are action-reaction force pairs (a common mistake). In this case, the normal force and weight need to be equal in magnitude to explain why there is no upward acceleration of the object. For example, a ball that bounces upwards accelerates upwards because the normal force acting on the ball is larger in magnitude than the weight of the ball.<br>question: is the normal force equal to the force of gravity?<br>answer: |
| Target Completion → | yes |

# WIC (WORD IN CONTEXT) TASK

| | |
|---|---|
| Context → | An outfitter provided everything needed for the safari.<br>Before his first walking holiday, he went to a specialist outfitter to buy some boots.<br>question: Is the word ‘outfitter’ used in the same way in the two sentences above?<br>answer: |
| Target Completion → | no |

# COPA TASK

| | |
|---|---|
| Context → | My body cast a shadow over the grass because |
| Correct Answer → | the sun was rising. |
| Incorrect Answer → | the grass was cut. |

# WSC (WINOGRAD SCHEMA CHALLENGE) TASK

| | |
|---|---|
| Context → | Final Exam with Answer Key<br>Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in *bold* refers to.<br>=====<br>Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires *his* financial support.<br>Question: In the passage above, what does the pronoun "*his*" refer to?<br>Answer: |
| Target Completion → | mr. moncrieff |

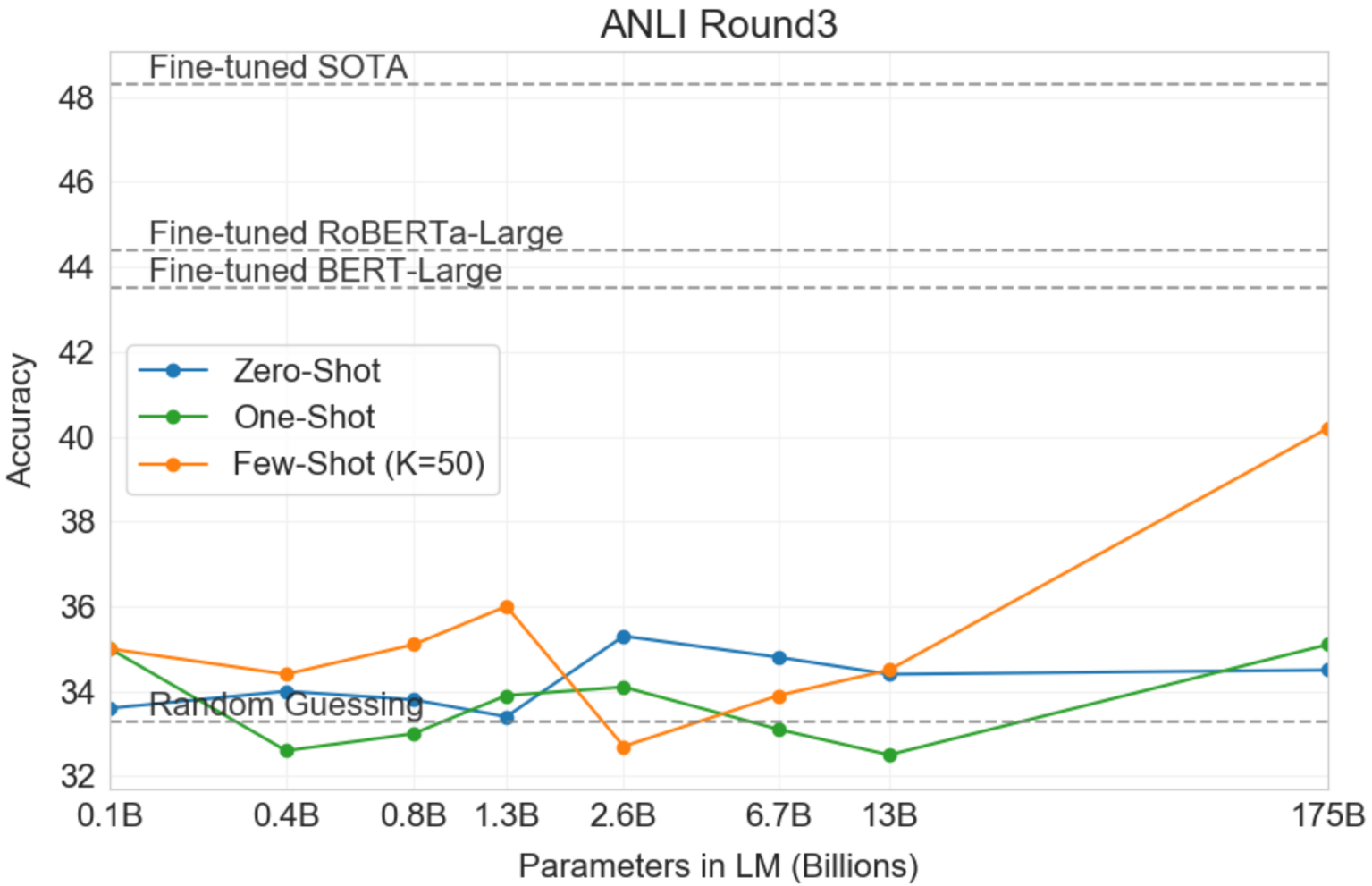
# RECORD TASK

| | |
|---|---|
| Context → | (CNN) Yuval Rabin, whose father, Yitzhak Rabin, was assassinated while serving as Prime Minister of Israel, criticized Donald Trump for appealing to "Second Amendment people" in a speech and warned that the words that politicians use can incite violence and undermine democracy. "Trump's words are an incitement to the type of political violence that touched me personally," Rabin wrote in USAToday. He said that Trump's appeal to "Second Amendment people" to stop Hillary Clinton -- comments that were criticized as a call for violence against Clinton, something Trump denied -- "were a new level of ugliness in an ugly campaign season."<br><br>- The son of a former Israeli Prime Minister who was assassinated wrote an op ed about the consequence of violent political rhetoric.<br>- Warns of "parallels" between Israel of the 1990s and the U.S. today. |
| Correct Answer → | - Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Donald Trump's aggressive rhetoric. |
| Correct Answer → | - Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Trump's aggressive rhetoric. |
| Incorrect Answer → | - Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Hillary Clinton's aggressive rhetoric. |
| Incorrect Answer → | - Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned U.S.'s aggressive rhetoric. |
| Incorrect Answer → | - Referencing his father, who was shot and killed by an extremist amid political tension in Israel in 1995, Rabin condemned Yitzhak Rabin's aggressive rhetoric. |

# ANLI TASK

| | |
|---|---|
| Context → | anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: $43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China.<br>Question: The loophole is now gone True, False, or Neither? |
| Correct Answer → | False |
| Incorrect Answer → | True |
| Incorrect Answer → | Neither |

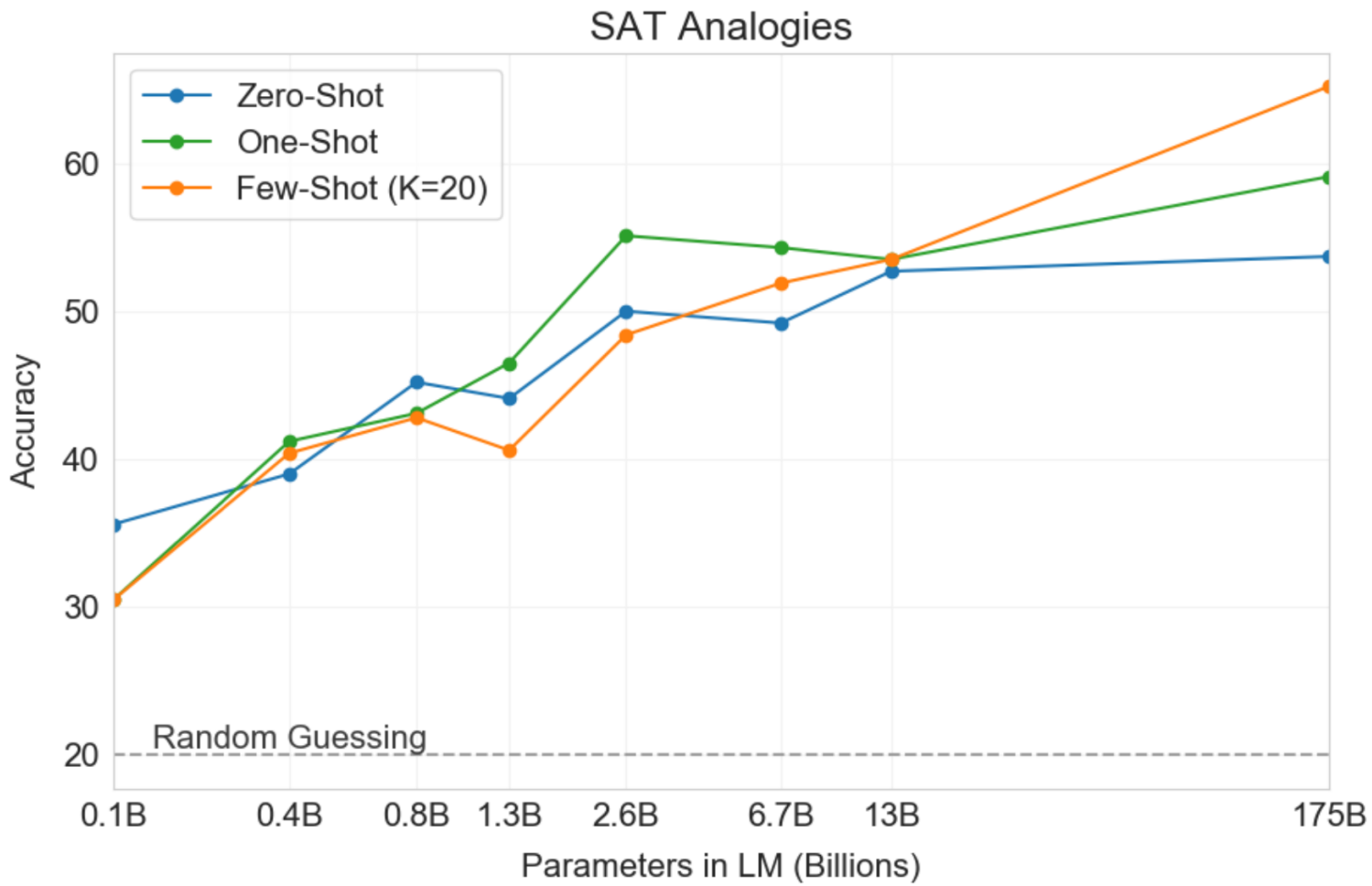
# PERFORMANCE ON ANLI TASK



# SAT ANALOGIES TASK

| | |
|---|---|
| Context → | lull is to trust as |
| Correct Answer → | cajole is to compliance |
| Incorrect Answer → | balk is to fortitude |
| Incorrect Answer → | betray is to loyalty |
| Incorrect Answer → | hinder is to destination |
| Incorrect Answer → | soothe is to passion |

# PERFORMANCE ON SAT ANALOGIES



# GPT3 CAN CORRECT GRAMMAR

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
**Good English output: I'd be more than happy to work with you on another project.**

---

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
**Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.**

---

Poor English input: The patient was died.
**Good English output: The patient died.**

---

Poor English input: We think that Leslie likes ourselves.
**Good English output: We think that Leslie likes us.**

---

Poor English input: Janet broke Bill on the finger.
**Good English output: Janet broke Bill's finger.**

---

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
**Good English output: Mary arranged for John to rent a house in St. Louis.**

---

# GENERATION OF NEWS ARTICLES

# HUMANS CANNOT DISTINGUISH HUMAN GENERATED VS GPT-3 GENERATED

| | Mean accuracy | 95% Confidence Interval (low, hi) | $t$ compared to control ($p$-value) | "I don't know" assignments |
|---|---|---|---|---|
| Control (deliberately bad model) | 86% | 83%–90% | - | 3.6 % |
| GPT-3 Small | 76% | 72%–80% | 3.9 (2$e$-4) | 4.9% |
| GPT-3 Medium | 61% | 58%–65% | 10.3 (7$e$-21) | 6.0% |
| GPT-3 Large | 68% | 64%–72% | 7.3 (3$e$-11) | 8.7% |
| GPT-3 XL | 62% | 59%–65% | 10.7 (1$e$-19) | 7.5% |
| GPT-3 2.7B | 62% | 58%–65% | 10.4 (5$e$-19) | 7.1% |
| GPT-3 6.7B | 60% | 56%–63% | 11.2 (3$e$-21) | 6.2% |
| GPT-3 13B | 55% | 52%–58% | 15.3 (1$e$-32) | 7.1% |
| GPT-3 175B | 52% | 49%–54% | 16.9 (1$e$-34) | 7.8% |

# HARD TO IDENTIFY AS FAKE

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

# EASIER TO IDENTIFY AS FAKE

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

# SUMMARY

- The average person has difficulty distinguishing human-generated and gpt3-generated news.
- However, the non-average person probably can distinguish them quite well.
- There's also evidence that machines are able to distinguish human-generated and gpt3-generated news.
- This has great significance for preventing abuse of AI technology.