# GPT-3 Weighted Shortcomings

## Initial Experiments

### Model: text-davinci-002

#### Self-curated samples: Max 100, top\_p 1, fp 0, pp 0

##### Zero-Shot

###### Q/A Format

Q: A person is wearing a mask and shouting in a bank. Who it **cannot** be? Name three.

[Temp 0] A: A bank robber, a bank employee, or a customer.

[Temp 0.7] A: A: The President, a doctor, or a bank robber.

Q: A person is wearing a mask and shouting in a bank. Which person is he less likely to be? Name three.

[Temp 0] A: He is less likely to be a bank teller, a bank customer, or a bank robber.

[Temp 0.7] A: He is less likely to be a bank teller, a bank customer, or a bank robber.

Q: What is less likely to be on a bar table? Name three.

[Temp 0] A: A bar table is less likely to have a lamp, a vase, or a book.

[Temp 0.7] A: A lamp, a book, a phone.

Q: What is not likely at all to be on a bar table? Name three.

[Temp 0] A: A bar table is not likely to have a sink, a stove, or a fridge.

[Temp 0.7] A: A cat, a lamp, and a book.

Q: Someone gives another person a flower. What are the **less likely** reasons? Name three.

[Temp 0] A: The person is angry with the other person, the person is sad, the person is trying to be romantic.

[Temp 0.7] A: One possible reason someone might give another person a flower is to show their affection for that person. However, there are many other reasons someone might give another person a flower. Some other possible reasons include: to apologize, to show appreciation, to show sympathy, or to show congratulations.

Q: Someone gives another person a flower. What are the impossible reasons? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: [No Answer!]

Q: Someone gives another person a flower. What are the impossible intentions? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: [No Answer!]

Q: Someone gives another person a flower. What cannot be the correct reasons? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: The first person is not interested in flowers, the second person is not interested in receiving gifts, the flowers are not fresh.

Q: Someone waits in a medical office's waiting room. What is the less likely reason to wait? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: [No Answer!]

Q: Someone waits in a medical office's waiting room. What are the less likely reasons to wait? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: [No Answer!]

Q: Someone waits in a medical office's waiting room. What is the person not waiting for? Name three.

[Temp 0] A: The person is not waiting for an appointment, the doctor, or the nurse.

[Temp 0.7] A: The person is not waiting for a doctor, a nurse, or a receptionist.

Q: Someone waits in a medical office's waiting room. What is the person less likely waiting for? Name three.

[Temp 0] A: The person is less likely waiting for a haircut, a manicure, or a pedicure.

[Temp 0.7] A: The person is less likely waiting for an MRI, a Pap smear, or a blood test.

Q: Someone is sitting on a chair in a pedicure salon. What is the person less likely expecting? Name three.

[Temp 0] A: The person is less likely expecting the chair to be uncomfortable, the pedicure to be painful, or the salon to be dirty.

[Temp 0.7] A: The person is less likely expecting the chair to be comfortable, the pedicure to be relaxing, and the nails to be painted well.

###### Direct Completion

##### Few-Shot

###### Q/A Format

Dark green shows the few-shot prompts. Blue is the question. Answer is in white and when incorrect in red.

Q: John visits a zoo. What is John not expecting? Name three.

A: Zoo is a place to see animals. Therefore, John is not expecting seeing humans in cells, dead animals, or sad animals.

Q: Sam goes to new year's fireworks. What does Sam not expect? Name three.

A: New year's fireworks is a fun event, where people gather to watch fireworks in the night sky. Sam does not expect a bombing, mass shooting, or quarrel.

Case 1: NOT case. Female.

Q: Cassandra waits in a medical office's waiting room. What is Cassandra not waiting for? Name three.

A: Cassandra is waiting to see the doctor. Therefore, she is not waiting for the nurse, the receptionist, or the janitor.

Case 2: NOT case. Male.

Q: James waits in a medical office's waiting room. What is James not waiting for? Name three.

A: A medical office's waiting room is a place where people wait to see the doctor. Therefore, James is not waiting for the doctor to diagnose him, the doctor to prescribe him medication, or the doctor to give him a surgery.

Case 3: For comparison, here are the positive relation results:

Q: James waits in a medical office's waiting room. What is James waiting for? Name three.

A: James is waiting to see the doctor, to get a medical procedure done, or to get test results.

###### Direct Format

- John visits a zoo. Zoo is a place to see animals. Therefore, John is not expecting seeing humans in cells, dead animals, or sad animals.

- Sam goes to new year's fireworks. New year's fireworks is a fun event, where people gather to watch fireworks in the night sky. Therefore, Sam does not expect a bombing, mass shooting, or quarrel.

Case I: NOT case

- James waits in a medical office's waiting room. James is not waiting for a doctor who is rude, a dirty waiting room, or long wait time.

Case II: Less likely case

- James waits in a medical office's waiting room. James is less likely waiting for

- a pleasant conversation with the receptionist

- to be seated next to a crying baby

- the doctor to give him bad news

Assuming that these people have no prior knowledge of what will happen, they would not expect any of the negative events to occur.

Case III: Probably

- James waits in a medical office's waiting room. James is probably waiting for an appointment with the doctor. Therefore, James does not expect to see a fight, a robbery, or a medical emergency.

Case IV: Definitely

- James waits in a medical office's waiting room. James is definitely waiting for a doctor, nurse, or medical staff member. Therefore, James does not expect the waiting room to be empty, or full of other patients.

2. What are some things that you would expect to see or experience when you visit a new country?

- Meeting new people from different cultures

- Trying new food

- Seeing new sights

- Learning about new customs

#### VisualComet Samples: Max 100, top\_p 1, fp 0, pp 0

##### Image ID: ? [Home Page]

Q: A person is holding onto a bronze statue while waves of water crash around him. What is less likely to have happened before? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: [No Answer!]

Q: A person is holding onto a bronze statue while waves of water crash around him. What is not his/her intention? Name three.

[Temp 0] A: [No Answer!]

[Temp 0.7] A: To keep the statue from being swept away, to stay dry, to keep the statue from being damaged.

Q: A person is holding onto a bronze statue while waves of water crash around him. What is less likely to happen after? Name three.

[Temp 0] A: The person is less likely to drown, less likely to be pulled away by the waves, and less likely to be hit by debris.

[Temp 0.7] A: The person is less likely to be pulled away by the waves, to be hit by debris, or to be struck by lightning.

##### Image ID: lsmdc\_0010\_Frau\_Ohne\_Gewissen/0010\_Frau\_Ohne\_Gewissen\_01.30.39.397-01.30.46.003@0.jpg

#### ATOMIC-2022 Samples: Max 100, top\_p 1, fp 0, pp 0

Table

Description automatically generatedTable

Description automatically generated

###### Social Interaction

Note: getting the car repaired is not clear.

**xWant**

Q: Harry gets John's car repaired. What harry did not want to do? Name three.

[t=0.0] A: Harry did not want to pay for the repairs, he did not want to do the repairs himself, and he did not want to have to ask John for help.

[t=0.7] A: Harry did not want to repair John's car, drive John's car, or work on John's car.

Q: Harry asks John was okay. What Harry did not want as a result? Name three.

[t=0.0] A: Harry did not want John to be angry, upset, or disappointed.

[t=0.7] A: Harry did not want John to be hurt, to be angry, or to be upset.

**xIntent**

Q: Harry gets John's car repaired. What was not Harry's intention? Name three.

[t=0.0] A: Harry did not intend to fix John's car. Harry did not intend to pay for John's car. Harry did not intend to drive John's car.

[t=0.7] A: Harry did not intend to fix John's car. He did not intend to pay for the repairs. He did not intend to drive the car himself.

Q: Harry answers John's question. What was not Harry's intention? Name three.

[t=0.0] A: Harry's intention was not to answer John's question.

[t=0.7] A: Harry's intention was not to deceive John, to make John feel foolish, or to make John feel uncomfortable.

Q: Harry asks John was okay. What was not Harry's intention? Name three.

[t=0.0] A: Harry's intention was not to ask John was okay.

[t=0.7] A: [No Answer!]

**xNeed**

Q: Harry gets John's car repaired. What Harry did not need beforehand? Name three.

[t=0.0] A: Harry did not need a wrench, a screwdriver, or a hammer.

Q: Harry asks John was okay. What Harry did not need before that? Name three.

[t=0.0] A: Harry did not need to ask John if he was okay, he needed to ask him for his name, address, and phone number.

[t=0.7] A: Harry did not need food, water, or shelter before that.

###### Event-Centered

**isBefore**

Q: Harry asks John was okay. What was less likely to happen before? Name three.

[t=0.0] A: [No Answer]

[t=0.7] A: [No Answer]

Q: Harry asks John was okay. What did not need to happen before? Name three.

[t=0.0] A: [No Answer]

[t=0.7] A: Harry did not need to talk to John, John did not need to be okay, and Harry did not need to ask.

**isAfter**

Q: Harry asks John was okay. What is not going to happen after? Name three.

[t=0.0] A: [No Answer]

[t=0.7] A: [No Answer]

Q: Harry asks John was okay. What is less likely to happen after? Name three.

[t=0.0] A: [No Answer]

[t=0.7] A: [No Answer]

**HinderedBy**

Q: Harry asks John was okay. What this even cannot be hindered by? Name three.

[t=0.0] A: [No Answer]

[t=0.7] A: [No Answer]

**SubEvents**

Q: Harry asks John was okay. What events or actions this does not include? Name three.

[t=0.0] A: This does not include John's thoughts, feelings, or opinions.

[t=0.7] A: [No Answer]

**Physical-Entity**

### Weight Comparison

Another method to look at weights in commonsensical relationships is through comparison, as it is hard to quantify weights to relationships. For example, saying maybe a relationship holds is rather difficult for humans to evaluate too. However, if we compare triples with each other in terms of weight, we can better get a good comparative understanding of relations. In the following examples, we show that LLMs lack this kind of weight comparison commonsense. Added to this, even curated commonsense KGs, such as ConceptNet lack this understanding as shown in the literature review. One way to evaluate LLMs and ConceptNet is through devising a test set to evaluate them. Improvement of LLMs can be through extra relevant training data, prompting technique or knowledge-based prompting, as in super-prompting.

#### Normal Q/A

Q: Is it more likely to see spoon close to a bowl or a fork?

A: It is more likely to see a spoon close to a bowl than a fork.

Q: Is it more likely to see spoon close to a bowl full of salad or a fork?

A: It is more likely to see a spoon close to a bowl full of salad than a fork.

Next try to make the question clear

Q: Is it more likely to see spoon close to a bowl or see fork close to a bowl? A: It is more likely to see a spoon close to a bowl than a fork close to a bowl. Q: Is it more likely to see spoon close to a salad bowl or see fork close to a salad bowl? A: It is more likely to see a spoon close to a salad bowl than a fork close to a salad bowl.

#### Chain-of Thought Prompting

Q: Is it more likely to see spoon close to a bowl or a fork?

A: Spoon is used for eating more liquid food, such as soup. Bowl usually contains soup. Therefore, it is more likely to see spoon close to a bowl.

Q: Is it more likely to see spoon close to a bowl full of salad or a fork?

A: It is more likely to see a spoon close to a bowl full of salad since salad generally has more dressing or vinaigrette than other types of food.

Second Run

A: It is more likely to see spoon close to a bowl full of salad.

Third Run

A: It is more likely to see a fork close to a bowl full of salad because salad is a dry food.

Fourth run

A: Fork is used for eating more solid food, such as salad. Bowl full of salad usually contains salad. Therefore, it is more likely to see spoon close to a bowl full of salad.

# Tasks

Some tasks can be interesting to analyze:

* Generating weighted completions (predicate), given a narrative (subject and predicate). The weights can be high (very plausible), mid (somewhat plausible), low (not plausible).
  + X repels Y’s attack 🡪 as a result X 🡪
    - [wants] X wants to file a police report - ATOMIC
    - [may want] X may want to call for help - Ours
    - [does not want] X does not want to be injured - Ours
* Discriminative task: Given a narrative (subject and object), predict which category of plausibility the result ends up in.
  + X repels Y’s attack 🡪 as a result X 🡪
    - [**wants**] - ATOMIC
      * Correct: X wants to file a police report
      * Incorrect: X wants to run away
    - [may want] X may want to call for help – Problematic a bit
      * Correct: X may want to call for help
      * Incorrect: X may want to sleep?
    - [**does not want**]
      * Correct: X does not want to be injured - annotation
      * Incorrect: to file a police report – We can reuse
* The same generative and discriminative tasks can be analyzed in specific contexts, e.g., in a medical office, in a library, in a restaurant, etc. The results could be different given the contexts.
  + Context: Inside a boxing ring; X repels Y’s attack 🡪 as a result X 🡪
    - [wants] to defend himself to not give points to the opponent
    - [may want]
    - [does not want]

# Possible Methods

* **Natural Language:** We can use narratives given in natural language format and predict the results in the same format and then convert to the KG style. The weights also represented in natural language format.
  + Zero-shot; Few-shot; Supervised.
* **Structured:** For comparison, we can have the triples represented in a structured format and the weights can be represented with special tokens, e.g. [LOW], [MID], and [HIGH].
  + Zero-shot; Few-shot; Supervised.

# Experiments

Based on our Symmetry paper, we could see that more than half of predicates analyzed had worse performance in the “less likely” scenario, such as holding, with, in, has, behind, and in front of. In case of “made of”, the accuracy ­of less likely was much more than more likely, which was an outlier. These experiments hint at some possible shortcomings of large language models (LLMs), when it comes to negated predicate weights. Added to this, further unpublished experiments also showed possible inferior results, when experimenting with mid-level weights of predicates.

Table

Description automatically generated

Chart, bar chart

Description automatically generated

Added to our previous experiments, ATOMIC-2020 paper also has one example of a negated predicates, which is “NotDesires”, which is negated form of “Desires”. An example mentioned is that “doctors” likely desire to “cure patients,” but do not desire “malpractice suit”. The question is how NotDesired results can be compared with Desired? Are the results reported per predicate type? Could not find the human evaluations in the repository so far, but the automatic evaluation results are here: <https://github.com/allenai/comet-atomic-2020/tree/master/system_eval>. We can run the evaluation and do it for specific predicates. TODO

## Initial ATOMIC-2020 Experiments

For the initial manual experiments, we use some samples from the ATOMIC-2020 dataset’s dev partition. The triples in the ATOMIC-2020 consist of a subject, a predicate, and an object. For example, a subject can be “PersonX applies sunscreen”, the predicate can be “xIntent”, and the relevant object to be “to protect their skin”. The triples are categorized semantically into one of these categories: Physical-Entity, Event-Centered, or Social-Interaction. As the name shows, the Physical-Entity category is about physical entities and commonsense of them. Some physical-entity sample predicates are AtLocation, or CapableOf. The other category encompasses triples about event commonsense, such as “X runs out of steam” 🡪 “Causes” 🡪 “takes a break”. Last category covers social interactions, such as “X votes for Y” 🡪 xIntent 🡪 “to give support”.

The goal is to generate an object in natural language given a verbalized subject and predicate. We consider the verbalized subject and predicate to be a premise (NLI task analogy) and the object to be a hypothesis. The annotator decides if the hypothesis entails the premise or not. The hypothesis can be a correct statement, but if it does not entail the premise, it is considered to be wrong (does not entail).

We first experiment with the largest pretrained GPT-3 model available at the time, which is text\_davinci\_002, with zero-shot configuration, temperature of 0.7, max length of 10 tokens, and top\_p of 1. The results are of the experiments are self-evaluated using the NLI analogy described in the previous paragraph. The experiments are located under this folder: experiments/initial\_self\_evaluated/text\_davinci\_002\_t\_0\_7\_ml\_10\_top\_p\_1/

* Zero shot:
  + Without context:
    - Negated event-based predicates: 43.75 % accuracy
    - Negated social-based predicates: 45.83 % accuracy
  + With context:
    - Regular Social-based predicates: 84.85 % accuracy
    - Negated social-based predicates: 53.33 % accuracy

Based on these preliminary results, it seems that negating predicates results in lower accuracy, while adding context results in improved accuracy of the hypotheses.

## ATOMIC-2020 NotDesire

We notice that the ATOMIC-2022 paper already has a predicate, which is a negated version of the Desires predicate: NotDesires. As we are interested in weighted commonsense analysis, finding reported results from NotDesires is interesting. The paper does not provide per predicate evaluation results of the generated objects. The only per predicate result is related to the human-annotated objects aggregated.

TODO:

* Look at the ATOMIC-2020 code
* Look at the Symbolic Distillation using GPT-3 paper as well

## Random Large-Scale Negated Object Prediction Evaluation using GPT-3

To better establish the initial ATOMIC-2020 zero-shot experiments, we design an experiment to systematically analyze the accuracy of prediction when predicates are negated. We use ATOMIC-2020’s test partition dataset. Similar to the paper, we segment the types of triples into three categories of physical entities, events, and social norms. The predicates under each segment are shown under [this section](#_ATOMIC-2022_Samples:_Max). We choose N random samples from each of the groups. Then generate objects, given the verbalized subjects and predicates. We then evaluate each of them. We first do self-evaluation ourselves. Then ask three annotators from Amazon mTurk to do the same. We must do due diligence in selection of the annotators. We then plot the accuracy of the negated predicates. We also report the accuracy per predicate. The verbalized form of the predicates are used from the ATOMIC-2020 paper and are adjusted to show negation.

Implementation decisions:

* Starting with 10 random samples per predicate with seed of 66.
* Negated predicates are shown with prepending *Not*, such as *NotAtLocation*.
* NotDesires is kept in the negated predicates.

Adjusted evaluation categories based on Symbolic Distillation paper:

* True: Always or quite often true.
* Sometimes: Sometimes is true or true for some people. -or- Likely true.
* False: False or farfetched, at best. -or- Unlikely to be true.
* Invalid: This assertion makes no sense (i.e., "what does this even mean?!"). Or the concept is not relevant to the description at all!
* Unfamiliar: Cannot make a fair evaluation. Unfamiliar with one or both phrase.

""" TODO:

1. Stop generation at . (But causes problem for a. b. cases)

2. Assign Invalid to duplicate generations. Is this the fair way to do it?

3. [DONE] Assign Invalid to empty generations.

4. [DONE] Assign Invalid to generations with \_\_\_\_.

5. [Stops at “. ”] In cases of a. b., etc. Only choose the element under a. (first one).

"""

TODO:

* Use individual names instead of PersonX and PersonY
* Evaluation of zero-shot results:
  + Normal
  + Negated
* Compare normal and negated accuracies: zero-shot
* Doing the same for few-shot:
  + Normal
  + Negated
* Compare the accuracies per predicate
* Introduce our won method to improve upon the negated results, such as knowledge-base method

POTENTIAL PROBLEM: We do not have the evaluation of the regular predicates. Do we really need them for comparison? Or we can show the progress of accuracy from zero-shot to few-shot to knowledge-enhanced? Here are the options to obtain the results per predicate:

* Ask the authors of ATOMIC-2020 for that.
* Evaluate automatically instead, but not sure if automatic evaluation is good for negated predicates.
* We evaluate them ourselves, but results in extra cost.

TODO:

* [DONE] What to do with predicates, such as isFilledBy or xAttr, where there is a blank space in the prompt (\_\_\_)?
  + For now, avoided all heads including the string “\_\_\_” which resulted in elimination of isFilledBy.
* We need to refine the prompts, e.g. Causes prompt is “Sometimes sickness causes”. The negated prompt should be something like: “Sickness does not cause” and not “Sometimes sickness does not cause”.

TODO: Change the temperature and plot a graph of accuracy vs. temperature.

TODO: As the physical entities is not well diverse, use our own IPMU dataset that instead of AtLocation is more fine-tuned to have relations, such as Above, Below, Besides, etc.

**TODO:**

* Have a specific name for the phenomenon that we see where the negated predicates have lower performance. We can possibly call it Normal Flow Bias, where if a normal flow of sentence is broken, the language model has a hard time logically responding.

**Submission TODO:**

* ArXiv does not seem to accept anonymous submissions. We can use OpenReview’s anonymous submission server: <https://openreview.net/group?id=OpenReview.net/Anonymous_Preprint>. Does ACL itself use OpenReview?

**TODO:**

* Calculate (for one or a few predicates), the effect of temperature at least as a hyperparameter on the accuracy of the results. Both negated and normal.

**Solutions for improvement:**

1. Re-phrase the question. Refer to the paper that talks about language modes already knowing. Re-ask the same question in a re-phrased manner, e.g., Really?! You think we can’t use hammer for cooking? Can you explain and decide?
2. Chain-of-thought prompting. What chain of thought? Should we go with the regular predicate (negated of negated predicate) and then try the negated one?
3. Extra automatic knowledge? What knowledge is missing?

TODO:

* Look at the wrong results and categorize the usual types of mistakes to help in diagnosis.

### mTurk Evaluation

#### Description

Based on your own commonsense, choose one of the five options. Examples are provided in the description.

IMPORTANT: Please note the CANNOT, DO Not, and other negated cases.

#### Instructions

**Unfamiliar to me to judge**

**Example:** PersonX discovers a new planet. The planet is in the Alpha Centauri system.

**First part and second part are not related! Or not enough information to judge.**

Example: PersonX rides a bike. Elephants are not birds. (Although second part is correct, it is not related to the first part)

**Makes sense:**

**Example:** It is NOT likely to see elephant on table.

**Does not make sense:**

**Example:** It is likely to see elephant on table.

#### UI

Graphical user interface, text, application

Description automatically generated

#### How many samples to evaluate?

|  |  |  |
| --- | --- | --- |
| **Paper** | **Human Evaluation Size** | **Notes** |
| COMET-ATOMIC-2020 | 3k random instances per KG. Three different knowledge graphs. | “To ensure high-quality annotations, we qualified a pool of 173 workers through a paid qualification task that tested their ability to follow directions and provide reasonable answers to the qualification test. The qualification test contained 6 manually selected tuples from ATOMIC and CONCEPTNET, including both easy and tricky relations to annotate. A worker was qualified if they provided 100% acceptable answers. Workers providing 5 of 6 correct answers were also accepted only when they provided a reason- able written substantiation for their incorrect choice.”  “Workers were presented with knowledge tuples in the form of (*head*, *relation*, *tail*) for an- notation. To expedite the human assessment of the tuples, each *relation* (e.g., xWant or AtLocation) was translated into a human-friendly natural language form (e.g., “as a result, PersonX wants” and “located or found at/in/on”, respectively; cf. Appendix ). The workers were asked to rate the tuples along a 4-point Likert scale: *always/often* – the knowledge assertion presented is always or often true, *some- times/likely* – it is sometimes or likely true, *farfetched/never* – it is false or farfetched at best, and *invalid* – it is invalid or makes no sense. Any tuples receiving the former two labels are ranked as Accept and latter two as Reject. The workers were also given a choice to opt out of assessment if the concepts were too unfamiliar for a fair evaluation (No Judgment). Each task (HIT) included 5 tuples of the same relation type, and each tuple was labeled by 3 workers. For the results, we take the majority vote among the 3 workers.” |
| CommonsenseEmergence-2021 | ? | “For the human evaluation (Accept % in Table 3), we use the procedure described in [Hwang et al., 2021].” |
| Commonsense QA 2.0 | In order of 1000s?! |  |
| Chain-of-Thought | Depends on the dataset:   * SayCan: 120 * GSM8k: 1,319 * SVAMP: 1000 * AQuA: 254 * MAWPS: 395-600 | * Standard deviation is reported for multiple runs with different random seeds. |
| I’m Not Mad | * Sample 100 events from ATOMIC. 5 evaluators each. Each event has 9 relations and 10 options. So total of 9,000 samples. |  |
| It's Not Rocket Science | Two types of evaluations:   * Absolute evaluations: 50 random narratives for each task. Two tasks of Idioms and Similes. * Comparative evaluation: 100 random narratives for idioms and 75 for similes. |  |
| Penguins Don’t Fly | I guess 1,500. As there are 300 generics and 5 top from each generic are annotated. |  |
| Symbolic Knowledge Distillation | * 3,000 before filtering, so at least 1,000 are there. |  |