

# WHEN DOES IN-CONTEXT LEARNING FALL SHORT AND WHY? A STUDY ON *Specification-Heavy* TASKS

## - APPENDIX -

**Anonymous authors**

Paper under double-blind review

### A EXPERIMENTAL DETAILS

#### A.1 PILOT EXPERIMENT

For ChatGPT, Davinci, and GPT-4, we use 8-shot demonstrations for all datasets except for 2-shot for DocRED and MAVEN-ERE (due to the limited input length), and 10-shot for FewRel 2.0 (the same setting as the previous few-shot learning). For the other LLMs, we can only use zero-shot demonstrations for DocRED and MAVEN-ERE due to the limited input length. For zero-shot experiments, we additionally incorporate an output format description in the prompt. For MATRES, MAVEN-Causal, MAVEN-Subevent, and MAVEN-Temporal, we require LLMs to output the relations for a single event pair at a time. Table 4 to 9 shows the input-output format of 6 representative tasks of each task type in the pilot experiments.

We query the APIs provided by OpenAI: `gpt-3.5-turbo`, `text-davinci-003`, and `gpt-4` for utilizing ChatGPT, Davinci, and GPT-4, respectively. The time period for our access to these APIs was from June 1 to June 30, 2023. All the API parameters are set to default. The FLAN-UL2, FLAN-T5, and Vicuna models are downloaded from Hugging Face Transformers (Wolf et al., 2020) and the model keys in Transformers are `google/flan-ul2`, `google/flan-t5-{size}`, and `lmsys/vicuna-7b-v1.1`, respectively. The Alpaca model is downloaded from its official GitHub repository<sup>1</sup>. We obtain the LLaMA’s checkpoint by applying to Facebook<sup>2</sup>. The releases of Alpaca and Vicuna are the model diffs to LLaMA and we manually merge them. We use greedy search as the decoding strategy for all the open-sourced models: FLAN-UL2, FLAN-T5, Alpaca, and Vicuna.

#### A.2 ANALYTICAL EXPERIMENTS

**Minor Modifications in Contexts** We conduct minor modifications in the contexts of 50 instances sampled from the *accurate* predictions. An example of minor modification in contexts for relation extraction is shown in Table 10. We aim to minimize the number of modified words while ensuring the modification changes the gold label.

**Exploring whether LLMs Ignore all the Contexts** To investigate whether LLMs ignore all the contexts or some specific words for the 27 instances with unchanged predictions, we query GPT-4 with only the questions while excluding the contexts. Take relation extraction as an example, as shown in Table 11, we omit the contexts and directly query GPT-4 the relationships of two entities with only entity names: “*What is the relationship between Wen Qiang and bribes?*”. We observe that in 18 cases (66.7%), LLMs predict the same with and without contexts, which suggests that LLMs ignore all the contexts and give predictions based on their parametric knowledge. For the other cases (33.3%), LLMs utilize the contexts for predictions but neglect specific words.

**Performance Degradation as Context Length Increases** Figure 1 shows the full curve of performance degradation as context length increases on the DocRED task.

<sup>1</sup>[https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca)

<sup>2</sup><https://github.com/facebookresearch/llama>

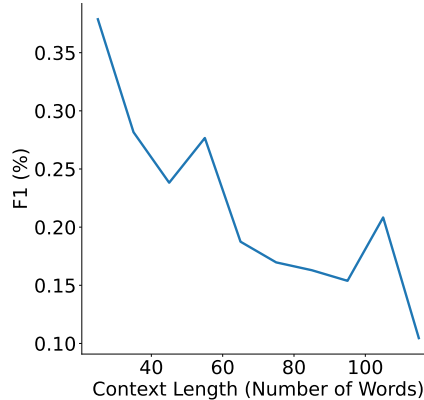


Figure 1: F1 scores (%) decrease as context length, i.e., the number of words, increases.

FewNERD	TACRED	ACE 2005 (ED)	ACE 2005 (EAE)	GoEmotions
67.8	73.4	74.0	75.8	52.9
+0.4	+0.7	+3.5	+1.5	-2.2

Table 1: F1 scores (%) and improvements of continual fine-tuning on instruction-tuned FLAN-UL2. The improvements are compared to fine-tuning results.

**Extensive Prompts with Schema Descriptions** We employ detailed descriptions for predefined schemata in prompts. Specifically, we sample 100 instances each from FewNERD, TACRED, ACE 2005 (ED), ACE 2005 (EAE), and GoEmotions tasks for evaluation. Table 12 shows the schema descriptions for GoEmotions.

### A.3 FINE-TUNING

We fine-tune FLAN-T5 (Small, Base, Large, XL, and XXL) and FLAN-UL2 on investigated tasks as text generation tasks. The input and output formats of fine-tuning are the same as LLM+ICL in pilot experiments except that instructions and demonstrations are excluded from input. The hyperparameters for fine-tuning are shown in Table 2. The experimental results of all fine-tuned models on all the investigated specification-heavy tasks are shown in Table 3. All the fine-tuning experiments are conducted on Nvidia A100 GPUs, which approximately consume 1,000 GPU hours.

### A.4 INSTRUCTION TUNING

We instruction-tune FLAN-UL2 on mixed datasets. The mixed datasets are sampled from original training sets and we sample up to 30k instances per dataset to balance the sizes of datasets. We utilize the examples-proportional mixing method (Raffel et al., 2020) in the training process and the mixing rate is 3k. We instruction-tune FLAN-UL2 with 32k gradient steps using AdamW (Loshchilov & Hutter, 2018) optimizer. The batch size is 32 and the learning rate is  $1.0 \times 10^{-3}$ . The maximum input and output sequence lengths are all 512. The instruction tuning experiments are conducted on Nvidia A100 GPUs, which consume about 1,200 GPU hours.

### A.5 CONTINUAL FINE-TUNING ON INSTRUCTION-TUNED FLAN-UL2

The effectiveness of fine-tuning could potentially be enhanced through existing techniques, such as two-stage fine-tuning (Phang et al., 2018; Li et al., 2019; Garg et al., 2020; Qiu et al., 2020). We follow this method to continue training on instruction-tuned FLAN-UL2 for individual tasks. We conduct experiments on FewNERD, TACRED, ACE 2005 (ED), ACE 2005 (EAE), and GoEmotions tasks and find that the results are consistently improved, which is shown in Table 1.

	FLAN-T5 <sub>SMALL</sub>	FLAN-T5 <sub>BASE</sub>	FLAN-T5 <sub>LARGE</sub>	FLAN-T5 <sub>XL</sub>	FLAN-T5 <sub>XXL</sub>	FLAN-UL2
Epoch	10	10	10	6	6	6
Batch Size	256	128	32	32	16	32
Warmup Rate	0.1	0.1	0.1	0.1	0.1	0.1
Learning Rate	$3.0 \times 10^{-5}$	$3.0 \times 10^{-5}$	$1.0 \times 10^{-5}$	$1.0 \times 10^{-5}$	$5.0 \times 10^{-6}$	$1.0 \times 10^{-5}$

Table 2: Hyper-parameters used in fine-tuning experiments.

Task	FLAN-T5 <sub>SMALL</sub>	FLAN-T5 <sub>BASE</sub>	FLAN-T5 <sub>LARGE</sub>	FLAN-T5 <sub>XL</sub>	FLAN-T5 <sub>XXL</sub>	FLAN-UL2
CoNLL-2003	73.0	87.0	89.8	91.4	92.2	92.5
ACE-2005 (NER)	67.7	81.2	83.2	86.4	88.6	89.3
FewNERD	55.2	63.8	65.0	66.2	67.2	67.4
TACRED	64.2	70.0	70.3	73.0	73.2	72.7
SemEval	60.9	81.2	73.8	84.1	88.2	87.9
FewRel 2.0	60.3	72.7	72.5	73.1	74.6	74.2
DocRED	15.3	32.8	36.0	51.2	53.6	54.5
ACE 2005 (ED)	40.0	62.2	61.3	64.9	69.0	70.5
MAVEN	48.7	60.0	60.4	62.3	64.4	64.2
RichERE (ED)	35.3	50.2	49.4	54.0	57.1	61.3
ACE 2005 (EAE)	33.9	42.5	42.4	59.4	72.8	74.3
RichERE (EAE)	42.6	60.8	60.0	66.7	71.5	72.1
MATRES	6.6	24.8	12.5	36.3	31.4	35.9
MAVEN-Causal	7.9	16.2	19.7	28.9	27.5	27.7
MAVEN-SubEvent	13.6	13.5	18.3	22.1	18.2	22.9
MAVEN-Temporal	13.7	18.0	18.9	23.5	23.7	24.5
GoEmotions	43.0	50.9	48.3	51.6	54.1	55.1
SST-5	51.6	57.8	59.9	61.3	61.2	62.0

Table 3: Experimental results (%) of fine-tuned models on all the investigated tasks.

## REFERENCES

- Siddhant Garg, Thuy Vu, and Alessandro Moschitti. TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of AAAI*, pp. 7780–7788, 2020. URL <https://arxiv.org/pdf/1911.04118.pdf>.
- Zhongyang Li, Xiao Ding, and Ting Liu. Story ending prediction by transferable BERT. In *Proceedings of IJCAI*, pp. 1800–1806, 2019. URL <https://www.ijcai.org/Proceedings/2019/0249.pdf>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of ICLR*, 2018. URL <https://openreview.net/pdf?id=Bkg6RiCqY7>.
- Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018. URL <https://arxiv.org/pdf/1811.01088.pdf>.
- Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897, 2020. URL <https://arxiv.org/abs/2003.08271>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. URL <https://jmlr.org/papers/volume21/20-074/20-074.pdf>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.

**Instruction**

Please recognize entities for the given text and classify them into a suitable type. The collection of types is as follows: art-broadcastprogram, art-film, art-music, art-other, art-painting, art-writtenart, building-airport, building-hospital, building-hotel, building-library, building-other, building-restaurant, building-sportsfacility, building-theater, event-attack/battle/war/militaryconflict, event-disaster, event-election, event-other, event-protest, event-sportsevent, location-bodiesofwater, location-GPE, location-island, location-mountain, location-other, location-park, location-road/railway/highway/transit, organization-company, organization-education, organization-government/governmentagency, organization-media/newspaper, organization-other, organization-politicalparty, organization-religion, organization-showorganization, organization-sportsleague, organization-sportsteam, other-astronomything, other-award, other-biologything, other-chemicalthing, other-currency, other-disease, other-educationaldegree, other-god, other-language, other-law, other-livingthing, other-medical, person-actor, person-artist/author, person-athlete, person-director, person-other, person-politician, person-scholar, person-soldier, product-airplane, product-car, product-food, product-game, product-other, product-ship, product-software, product-train, product-weapon.

**Demonstrations**

Text: Peter Nicol Russell's company, P. N. Russell and Company, constructed the heritage listed Denison Bridge at Bathurst.

Answer: Peter Nicol Russell: person-other; P. N. Russell and Company: organization-company; Denison Bridge: building-other; Bathurst: location-GPE;

<other demonstrations >

**Query**

Text: In 1926, before making his first-class debut, he played two matches for the Marylebone Cricket Club ( MCC ) against Ireland in Belfast and Dublin.

Answer:

**Answer**

Marylebone Cricket Club: organization-sportsteam; MCC: organization-sportsteam; Belfast: location-GPE; Dublin: location-GPE;

Table 4: An input-output example of FewNERD. The parts **Instruction**, **Demonstrations**, and **Query** are concatenated as the input to LLMs. **Answer** shows the output format.

**Instruction**

Please classify relationships between the two entities (marked with <entity> and </entity>). If the two entities have no relationships, please answer NA. Note the relation needs to be in the predefined set of relations. The set of relationships is as follows: org:founded, org:subsidiaries, per:date\_of\_birth, per:cause\_of\_death, per:age, per:stateorprovince\_of\_birth, per:countries\_of\_residence, per:country\_of\_birth, per:stateorprovinces\_of\_residence, org:website, per:cities\_of\_residence, per:parents, per:employee\_of, per:city\_of\_birth, org:parents, org:political/religious\_affiliation, per:schools\_attended, per:country\_of\_death, per:children, org:top\_members/employees, per:date\_of\_death, org:members, org:alternate\_names, per:religion, org:member\_of, org:city\_of\_headquarters, per:origin, org:shareholders, per:charges, per:title, org:number\_of\_employees/members, org:dissolved, org:country\_of\_headquarters, per:alternate\_names, per:siblings, org:stateorprovince\_of\_headquarters, per:spouse, per:other\_family, per:city\_of\_death, per:stateorprovince\_of\_death, org:founded\_by.

**Demonstrations**

Text: Named for the chief theorist of modern Zionism, Theodor Herzl, <entity> Kollek </entity> was born in Nagyvaszony near Budapest in 1911 and raised in <entity> Vienna </entity>.

Answer: (Kollek; per:cities\_of\_residence; Vienna)

<other demonstrations>

**Query**

Text: This news comes from Karr Ingham, an economist who created the <entity> Texas Petro Index </entity> (TPI), which is a service of the <entity> Texas Alliance of Energy Producers </entity>.

Answer:

**Answer**

(Texas Petro Index; org:parents; Texas Alliance of Energy Producers)

Table 5: An input-output example of TACRED.

**Instruction**

Please identify the events in the text and classify them into appropriate categories; The collection of categories is as follows: Becoming, Creating, Process\_start, Name\_conferral, Presence, Departing, Recording, Reporting, Cause\_to\_make\_progress, Bodily\_harm, Arranging, Supply, Getting, Change, Hold, Come\_together, Destroying, Hostile\_encounter, Cause\_change\_of\_position\_on\_a\_scale, Conquering, Expressing\_publicly, Killing, Dispersal, Agree\_or\_refuse\_to\_act, Coming\_to\_be, Communication, Bringing, Achieve, Sending, Change\_event\_time, Competition, Catastrophe, Attack, Surrounding, Military\_operation, Change\_sentiment, Award, Response, Commitment, Control, Process\_end, Motion, Deciding, Change\_of\_leadership, Earnings\_and\_losses, Choosing, Sign\_agreement, Placing, Death, Besieging, Escaping, Motion\_directional, Receiving, Self\_motion, Cause\_to\_amalgamate, Defending, Cause\_to\_be\_included, Removing, Statement, Preventing\_or\_letting, Openness, Causation, GiveUp, Expend\_resource, Aiming, Education\_teaching

**Demonstrations**

Text: The 2008 event also included a MySpace bus, where secret sets and interviews with bands took place.

Answer: included: Cause\_to\_be\_included; took place: Process\_start

<other demonstrations>

**Query**

Text: The ruling National Command of the Arab Socialist Ba'ath Party were removed from power by a union of the party's Military Committee and the Regional Command, under the leadership of Salah Jadid.

Answer:

**Answer**

removed: Removing

Table 6: An input-output example of MAVEN.

**Instruction**

Please extract event arguments and their roles for the events marked with <event> and </event> in the text, the possible roles must be chosen from the Roleset. If there are no roles for the event, place output "NA".

**Demonstrations**

Text: Schrenko is a scumbag and, although she won't get any actual jail time, I hope the fine she pleads to is at least as much as what she <event> stole </event>.

Role Set: Beneficiary, Time-Within, Money, Time-Starting, Time-Before, Place, Recipient, Giver, Time-Holds

Answer: Schrenko: Recipient

<other demonstrations>

**Query**

Text: Earlier documents in the case have included embarrassing details about perks Welch received as part of his <event> retirement </event> package from GE at a time when corporate scandals were sparking outrage.

Role Set: Time-Within, Time-Starting, Time-After, Time-Before, Time-Ending, Place, Entity, Position, Person, Time-Holds

Answer:

**Answer**

GE: Entity; Welch: Person

Table 7: An input-output example of the ACE 2005 event argument extraction task.

**Instruction**

Please classify the relation between two events/Timex in a given document. There are 10 types of relations: [before, overlap, contains, simultaneous, begins-on, ends-on, cause, precondition, subevent, and coreference]. In each document, 2 events/Timex are marked as `<event> event_name </event>` or `<Timex> Timex_name </Timex>`. If there is a relation type or multiple relation types, the answer form is Answer: [relation type 1, relation type 2,...].

**Demonstrations**

Text: The 2012 Great Britain and Ireland floods were a series of weather events that affected parts of Great Britain and Ireland periodically during the course of 2012 and on through the winter into 2013. The beginning of 2012 saw much of the United Kingdom experiencing droughts and a heat wave in March. A series of low pressure systems steered by the jet stream brought the wettest April in 100 years, and `<event> flooding </event>` across Britain and Ireland. Continuing through May and leading to the wettest beginning to June in 150 years, with flooding and extreme events occurring periodically throughout Britain and parts of Atlantic Europe. On 27 and 28 June and again on 7 July heavy rain events occurred from powerful thunderstorms that gathered strength as they travelled across mainland Britain. Severe weather warnings and a number of flood alerts were issued by the UK's Environment Agency, and many areas were hit by flash floods that overwhelmed properties and caused power cuts. A motorist was killed after his vehicle was caught by floodwater and landslides halted rail services between England and Scotland. The thunderstorms were the product of two fronts that collided over the British Isles 2013 warm air travelling from the Azores and cold water-laden air from the west. The second batch of flooding struck the South-West of England during the afternoon of 7 July, forcing the Met Office to issue its highest alert, Red (Take Action), due to the significant amounts of rainfall caused by a system travelling from Southern Europe, along with the warm, humid air the United Kingdom had seen in the run-up to the floods, which, like the June floods, caused thunderstorms. During the Autumn the most intense September low since 1981 brought widespread flooding and wind damage to the UK. Widespread flooding occurred again in November, December and `<Timex> January 2013 </Timex>` as more heavy rains overwhelmed the saturated ground.

The first event/ Timex : `<event> flooding </event>`

The second event/ Timex : `<Timex> January 2013 </Timex>`

Answer: [before]

`<other demonstrations>`

**Query**

Text: Tropical stormrumbia, known in the philippines as tropical storm unk, brought deadly flooding to the central and southern philippines in november and december 2000. The last of three consecutive tropical cyclones of at least tropical storm intensity to strike the Philippines, Rumbia `<Event> began </Event>` as a tropical depression on November 27, gradually intensifying to reach tropical storm intensity the next day. Strengthening later stagnated, and Rumbia would weaken back to depression status as it made landfall on the central Philippines on November 30. though the japan meteorological agency determined rumbia to have `<Event> dissipated </Event>` on december 2, the joint typhoon warning center continued to monitor the system over the next few days as it tracked across the south china sea. For a period of time beginning on December 5, Rumbia reorganized and strengthened back to tropical storm intensity before wind shear began to weaken the system. Located south of Vietnam on December 7, the storm's circulation center became devoid of convection, and by then Rumbia was declared by the JTWC to have dissipated. In the Philippines, Rumbia caused roughly US \$1 million in damage and 48 fatalities. Several transportation routes were suspended in the lead-up to the storm's landfall. As a result of the tropical storm, power outages occurred, especially in Surigao. Several towns and villages were flooding, displacing around 70,000 people and putting 4,100 people into temporary emergency sheltering.

The first event/ Timex : `<Event> began </Event>`

The second event/ Timex : `<Event> dissipated </Event>`

Answer:

**Answer**

[before]

Table 8: An input-output example of MAVEN-ERE.

<b>Instruction</b>
Please comprehend the emotions expressed in the given text. The set of emotions is as follows: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.
<b>Demonstrations</b>
Text: This is amazing man. Congrats and keep em coming
Answer: admiration; gratitude
<other demonstrations>
<b>Query</b>
Text: Help, help, I'm being repressed!
Answer:
<b>Answer</b>
<i>fear</i>

Table 9: An input-output example of GoEmotions.

<b>Original</b>
Text: Judy Gross' husband <entity> Alan Gross </entity> <b>was arrested</b> at the <entity> Havana </entity> airport in December 2009.
Answer: (Alan Gross; per:cities_of_residence; Havana)
<b>Modified</b>
Text: Judy Gross' husband <entity> Alan Gross </entity> <b>arrived</b> at the <entity> Havana </entity> airport in December 2009.
Answer: NA
Model output: (Alan Gross; per:cities_of_residence; Havana)

Table 10: An example of minor modifications in contexts for relation extraction. In this example, we change the phrase “**was arrested**” to “**arrived**”, and the golden label is also changed to NA.

<b>Instruction</b>
Please classify relationships between the two entities in the input. The input format is: (entity1; entity2). Answer the relation type only. You should classify the relationships based on your knowledge about the entities. If the two entities have no relationships, please answer NA. The set of relationships is as follows: org:founded, org:subsidiaries, per:date_of_birth, per:cause_of_death, per:age, per:stateorprovince_of_birth, per:countries_of_residence, per:country_of_birth, per:stateorprovinces_of_residence, org:website, per:cities_of_residence, per:parents, per:employee_of, per:city_of_birth, org:parents, org:political/religious_affiliation, per:schools_attended, per:country_of_death, per:children, org:top_members/employees, per:date_of_death, org:members, org:alternate_names, per:religion, org:member_of, org:city_of_headquarters, per:origin, org:shareholders, per:charges, per:title, org:number_of_employees/members, org:dissolved, org:country_of_headquarters, per:alternate_names, per:siblings, org:stateorprovince_of_headquarters, per:spouse, per:other_family, per:city_of_death, per:stateorprovince_of_death, org:founded_by.
<b>Query</b>
What is the relationship between Wen Qiang and bribes?
<b>Answer</b>
<i>per:charges</i>
<b>Model Prediction</b>
<i>per:charges</i>

Table 11: An example of TACRED omitting all the contexts. We evaluate LLMs using **Instruction** and **Query**, which contains only questions.

---

**Instruction**

Please comprehend the emotions expressed in the given text. The set of emotions is as follows: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. You should make the decision based on the definition of each type which is given as follows.

*admiration*: Finding something impressive or worthy of respect.

*amusement*: Finding something funny or being entertained.

*anger*: A strong feeling of displeasure or antagonism.

*annoyance*: Mild anger, and irritation.

*approval*: Having or expressing a favorable opinion.

*caring*: Displaying kindness and concern for others.

*confusion*: Lack of understanding, uncertainty.

*curiosity*: A strong desire to know or learn something.

*desire*: A strong feeling of wanting something or wishing for something to happen.

*disappointment*: Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.

*disapproval*: Having or expressing an unfavorable opinion.

*disgust*: Revulsion or strong disapproval aroused by something unpleasant or offensive.

*embarrassment*: Self-consciousness, shame, or awkwardness.

*excitement*: Feeling of great enthusiasm and eagerness.

*fear*: Being afraid or worried.

*gratitude*: A feeling of thankfulness and appreciation.

*grief*: Intense sorrow, especially caused by someone's death.

*joy*: A feeling of pleasure and happiness.

*love*: A strong positive emotion of regard and affection.

*nervousness*: Apprehension, worry, anxiety.

*optimism*: Hopefulness and confidence about the future or the success of something.

*pride*: Pleasure or satisfaction due to one's own achievements or the achievements of those with whom one is closely associated.

*realization*: Becoming aware of something.

*relief*: Reassurance and relaxation following release from anxiety or distress.

*remorse*: Regret or guilty feeling.

*sadness*: Emotional pain, sorrow.

*surprise*: Feeling astonished, startled by something unexpected.

*neutral*: Neither positive, negative, nor others.

---

**Demonstrations**

Text: This is amazing man. Congrats and keep em coming

Answer: admiration; gratitude

<other demonstrations>

---

**Query**

Text: Help, help, I'm being repressed!

Answer:

---

**Answer**

*fear*

---

Table 12: Detailed descriptions of each emotion option in the GoEmotions task.