Can large language models generate salient negative statements?

Hiba Arnaout[1], Simon Razniewski[2]

[1] *Max Planck Institute for Informatics, Germany*
[2] *Bosch Center for AI, Germany*

Abstract

We examine the ability of large language models (LLMs) to generate *salient* (interesting) *negative* statements about real-world entities; an emerging research topic of the last few years. We probe the LLMs using zero- and k-shot unconstrained probes, and compare with traditional methods for negation generation, i.e., pattern-based textual extractions and knowledge-graph-based inferences, as well as crowdsourced gold statements. We measure the correctness and salience of the generated lists about subjects from different domains. Our evaluation shows that guided probes do in fact improve the quality of generated negatives, compared to the zero-shot variant. Nevertheless, using both prompts, LLMs still struggle with the notion of factuality of negatives, frequently generating many ambiguous statements, or statements with negative keywords but a positive meaning.

1. Introduction

Motivation and Problem. Structured (knowledge graphs), and unstructured (text corpora) information are the backbone of many AI applications, such as question answering and chat bots. They mainly focus on storing positive knowledge, and mostly contain little negative knowledge. The open-world assumption, which advises to abstain from taking a stance on the truth of absent information, compromises the usability of both forms of machine knowledge. For instance, it is often the case that the *NBA*'s Basketball stars take a coaching position after their retirement. Notably, this is not *true* for *michael jordan.* Mining these surprising statements are useful to overcome limitations of applications like question answering systems. For example, querying Bing Chat[1] whether *michael jordan* invested in his team, *the chicago bulls*, returns an irrelevant answer about his achievements with the team. In fact, it is an interesting piece of information, that, even though he has a business-oriented mind, he did not monetarily invest in the *bulls*, but in other sports franchise, including an investment in the not so well-known team *the charlotte hornets.*

State of the Art. A new research area has emerged in the last few years, suggesting the importance of the explicit materialization of *important negative statements* about real-world subjects [1]. Several methodologies have been proposed [2, 3, 4, 5, 6]. The goal is to compile

KBC-LM'23: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2023

harnaout@mpi-inf.mpg.de (H. Arnaout); simon.razniewski@de.bosch.com (S. Razniewski)
https://hibaarnaout.com (H. Arnaout); http://simonrazniewski.com/ (S. Razniewski)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.microsoft.com/en-us/edge/features/bing-chat

Model	*Top Negative Statements*
Text	*didn't make his high school team* ~~*doesn't have social media*~~
KG	*isn't a basketball coach* *didn't play as a power forward*
ChatGPT *0-shot*	*didn't invent basketball* *didn't only play basketball* (positive)
ChatGPT *k-shot*	*never played for a team outside the u.s.* *didn't play for the bulls exclusively* (positive)
Alpaca *0-shot*	*didn't play for chicago until 84* (positive) *didn't win a championship for the lakers*
Alpaca *k-shot*	*wasn't the youngest player in the nba* *didn't win an oscar*
Human	*didn't buy stakes in the chicago bulls* *never coached the chicago bulls*

Table 1
Negative statements about *michael jordan* (salient , somewhat salient , nonsalient ; ~~incorrect~~), using different methodologies: text-based extractions, knowledge graph (KG) inferences, LLM generations, and human-written statements.

lists of statements (biographic summaries) about subjects, where the statements are truly negative, but also salient, unexpected, or normally mistaken as true positives. To compile these lists, different data sources and methodologies have been explored. In [2, 3], using web-scale knowledge graphs, candidate salient negatives are derived from existing positive statements about highly related entities. The computation relies on the local closed-world assumption, an assumption of completeness over identified relevant subgraphs, coupled with ranking metrics such as relative frequencies. Similarly, [4] explores graph embeddings to generate candidate negative statements, which are then scored using a fine-tuned language model (LM), by descending order of *negativity*. Textual sources have been explored in [5], where commonsense negative statements are extracted, by mining query logs, using pre-defined patterns. [6] makes use of the edit history of large collaborative encyclopedias, namely Wikipedia, by looking at sentences edited, where only an entity or a number are changed. The old version of the sentence is then considered an interesting negative statement.

LLMs for Negative Statements Generation. Recently, LMs have been examined about their ability to store factual knowledge about general topics [7, 8]. With LMs such as BERT [9], this was done via masked probing, e.g., "*Paris is the the capital of* `[MASK]`" generates *france* as the top prediction. With large LMs (LLMs), such as GPT-3 [10], autoregressive generation from textual prompts is the standard, e.g., "*Complete the following. Paris is..*", and receive the completion *the capital of France.* A few papers focused on the ability of these models to store and understand negative knowledge [11, 2, 12]. In [11], using masked probing, authors found that LMs, such as BERT, struggle to understand negation, predicting *fly* for the probe "*Birds cannot* `[MASK]`". In [2], methods to infer negative statements from knowledge graphs and text have been compared on a more specific negation task, namely generating *salient* negative *commonsense* statements. Results of these models are compared to ones using GPT-3. Even though performing better than BERT-like models [11], GPT-3 was not able to beat the SOTA model (inferences from KGs), neither on the true negativity of statements, nor their salience. More recently, [12] studies advanced LLMs, such as ChatGPT [13], on their ability to store negative knowledge in a constrained text generation and question answering tasks. The finding are contradictions in the LLM's belief, when comparing results of both tasks. For instance, LLMs generate the sentence "*Lions live in the ocean*", but answer "*No*" when asked "*Do lions live in the ocean?*". [12] is an important step towards examining LLMs' understanding of the falseness of statements, however, it has four main differences from our study: (i) our prompts are *not constrained* to commonsense knowledge; (ii) not constrained to puzzles around a set of words, but allowed to generate arbitrary subject-relevant statements; (iii) our comparison *includes SOTA baselines* from KG and text, not just LLMs; (iv) our study evaluates also the *salience* of outputs, not just their correctness.

We summarize our contributions as follows.

- We design *constraint-free prompts for LLM-based negation generation*, where we only instantiate the input subject.
- We examine LLMs' understanding of *salient factual negation*, finding that, even though they struggle with the notion of true negativity (-18% in correctness compared to SOTA model), on truly negative statements, the guided few-shot ChatGPT variant ranks first among models in salience.
- We study both *encyclopedic* and *commonsense* domains, finding that it is more challenging for LLMs to generate longer lists of salient *commonsense* negatives. For instance, the zero-shot ChatGPT variant shows a decrease of 22% in correctness@5 (compared to @1) for *commonsense* subjects. No decrease is observed for *encyclopedic* subjects.
- We compare the LLM-generated negative statements to existing SOTA methods, from text [5] and knowledge graphs [3].
- We measure the quality of the negative statements over two aspects, the correctness (true negativity) and salience (interestingness).

The data generated can be downloaded at: https://www.mpi-inf.mpg.de/fileadmin/inf/d5/research/negation_in_KBs/data.csv.

2. Probe Construction

Given a subject, we probe the LLM to generate a list of salient negative statements about it.

Zero-shot Probe. In this probe, we test the performance of the LLM without providing any samples in our instructions.

> *Write a list of* [n] *salient factual negated statement about* [SUBJECT].

The goal is to inspect the model's interpretation of the notion of *salient negation* without any prior examples nor definitions.

Guided Few-shot Probe. In this probe, we guide the model with both definitions and examples (for in-context learning).

> *A salient factual negated statement about an entity means that the statement doesn't hold in reality. Moreover, the negated statement is either surprising, unexpected, or useful to the reader. For example:*
>
> [EXAMPLE1]
>
> [. . .]
>
> *Given this definition and examples, write a list of* [n] *salient factual negated statement about* [SUBJECT].

In the following sample, we show a 4-shot probe with 2 salient and 2 nonsalient samples about different types of subjects, and request 3 salient negative statements about *lebanon* (LLM=ChatGPT).

> *A salient factual negated statement about an entity means that the statement doesn't hold in reality. Moreover, the negated statement is either surprising, unexpected, or useful to the reader. For example:*
>
> `penguins can't fly.`
>
> `tom cruise never won an oscar.`
>
> On the other hand, the following examples are factual negated statements that are not salient:
>
> `penguins can't run for presidency.`
>
> `tom cruise never won the nba best player award.`
>
> *Given this definition and examples, write a list of* 3 *salient factual negated statement about* `lebanon`.
>
> **Answer:**
>
> 1. *is not a desert country.*
> 2. *is not an oil-rich country.*
> 3. *is not a landlocked country.*

Model	cor@1	cor@3	cor@5	sal@1	sal@3	sal@5
	overall					
Text Extractions	0.38	0.30	0.33	0.63	0.69	0.68
KG Inferences	**0.94**	**0.76**	**0.75**	0.88	**0.84**	**0.83**
ChatGPT *0-shot*	0.71	0.65	0.60	0.73	0.73	0.71
ChatGPT *k-shot*	0.76	0.69	0.66	**0.89**	0.76	0.75
Alpaca *0-shot*	0.34	0.32	0.36	0.62	0.71	0.65
Alpaca *k-shot*	0.50	0.47	0.47	0.66	0.55	0.56
Human	*0.77*	*0.71*	*0.69*	*0.73*	*0.70*	*0.70*
	encyclopedic subjects					
Text Extractions	0.32	0.26	0.29	0.86	**0.91**	**0.88**
KG Inferences	**0.88**	**0.87**	**0.86**	**0.91**	0.86	0.83
ChatGPT *0-shot*	0.71	0.76	0.71	0.65	0.65	0.62
ChatGPT *k-shot*	0.76	0.73	0.74	0.89	0.74	0.72
Alpaca *0-shot*	0.32	0.33	0.38	0.63	0.70	0.64
Alpaca *k-shot*	0.52	0.45	0.48	0.69	0.59	0.58
Human	*0.78*	*0.70*	*0.69*	*0.69*	*0.64*	*0.65*
	commonsense subjects					
Text Extractions	0.47	0.36	0.39	0.44	0.47	0.48
KG Inferences	**1.0**	0.65	0.64	0.83	0.81	**0.83**
ChatGPT *0-shot*	0.72	0.55	0.50	0.81	**0.84**	**0.83**
ChatGPT *k-shot*	0.75	0.65	0.58	**0.89**	0.79	0.78
Alpaca *0-shot*	0.36	0.31	0.34	0.61	0.72	0.67
Alpaca *k-shot*	0.48	0.48	0.46	0.63	0.51	0.55
Human	*0.76*	***0.73***	***0.69***	*0.78*	*0.75*	*0.75*

Table 2
Results on correctness and salience of top negative statements (**best performance**, second best).

In Section 3, we experiment with different number of samples and different salient:nonsalient ratio (see Appendix D).

3. Evaluation

Data. We consider 50 subjects, 25 encyclopedic entities such as *elon musk*, and 25 commonsense concepts, such as *jogging* (Full list in Appendix A). Our intuition behind these choices is diversity: (i) in types, e.g., activities, occupations, people; and (ii) in popularity, e.g., *tom cruise* (a famous *hollywood actor*) and *peri gilpin* (a less known *tv actor*).

Methods. To compile lists of negative statements about these subjects, we consider:

- **Text Extractions**: The pattern-based method [5] relies on a handful of manually crafted patterns, in the form of *why-questions*, to extract interesting negative statements from rich query logs, e.g., "*why doesn't amazon.." with the completion "accept paypal*". We instantiate the query-log API with Google and Bing, merge the results, and rank by frequency.
- **KG Inferences**: The peer-based negation inference methodology [3] relies on a given KG to identify highly related entities to the input entity (called peers). Positive statements about these peers are used to infer candidate negatives, which are finally ranked using

Model	Correct	Incorrect	Ambiguous	Positive Meaning
Text Extractions	0.33	0.26	0.41	0
KG Inferences	0.75	0.13	0.12	0
ChatGPT *0-shot*	0.60	0.10	0.19	0.11
ChatGPT *k-shot*	0.66	0.17	0.10	0.07
Alpaca *0-shot*	0.36	0.42	0.13	0.09
Alpaca *k-shot*	0.47	0.38	0.04	0.10
Human	*0.69*	*0.05*	*0.12*	*0.14*
Sample Statement	*rabbits can't vomit*	*the beatles didn't tour*	*avocado isn't bad*	*lebanon isn't devoid of historical sites*

Table 3
Detailed look at the factuality and true negativity of generated statements.

statistical metrics, such as relative frequency, e.g., "*unlike similar physicists, such as max planck and albert einstein, stephen hawking never won the nobel prize in physics*". We instantiate the KGs to Wikidata [14] and Ascent [15], for encyclopedic/commonsense subjects, respectively.

- **ChatGPT** *0-shot*: The zero-shot probe introduced in Section 2 is submitted to ChatGPT [13] (May 2023 version).
- **ChatGPT** *k-shot*: The few-shot probe in Section 2, with k=3 (salient:nonsalient 3:0), is submitted to ChatGPT.
- **Alpaca** *0-shot*: The zero-shot probe introduced in Section 2 is submitted to Alpaca-13B, a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford [16].
- **Alpaca** *k-shot*: The few-shot probe from Section 2, with k=3 (salient:nonsalient 3:0), is submitted to Alpaca-13B.
  To ensure reproducibility, the randomness (temperature) for all LLMs variants is set to 0.
- **Human** [2]: We ask MTurkers to write lists of salient negative statements about a given subject. We show them examples of what a salient negative statement looks like. We collect, for each subject, two lists of statements from two workers. The performance is later measured as the average of the two.

Metrics. For the returned statements, we measure:

- **Correctness**: The true negativity (is it actually false?) and factuality of a statement (is it a judgeable statement?), e.g., not an opinion. We allow the labels: *correct*, *incorrect*, *ambiguous*, or *positive meaning*. Samples are shown in Table 3.
- **Salience**: The unexpectedness, informativeness, or interestingness of a statement. We allow: *salient* (1), *somehow salient* (0.5), and *nonsalient* (0).

Results are annotated on their salience by 2 domain-experts [3], with inter-annotator agreement = 60%. Correctness, the more straight forward metric of the two, was annotated by 1 of the domain-experts.

[2] We are aware of the risk that workers might use LLMs to generate these statements. In the absence of reliable detection tools on this newly emerging problem, we rely on our personal judgement as well as string matchings to discard untrustworthy answers. In particular, any response that matches the exact wording of one of the responses of the LLM baselines, or any near-duplicates in human-generations, were rejected.

[3] Experts on the topic of salient negative knowledge at web-scale.

k	sal:nonsal	Correctness	Salience
3	3:0	<u>0.72</u>	**0.54**
3	0:3	0.52	0.30
6	3:3	**0.80**	<u>0.40</u>
20	10:10	0.52	0.34

Table 4
Results given different values for the in-context learning parameters (**best performance**, <u>second best</u>).

True Factuality and Negativity of Statements. Results for correctness are shown in Table 2, and investigated further in Table 3. The *KG inferences* model ranks first on correctness overall. This is due to the factuality of KG statements. KG triples, especially encyclopedic ones, are expressed using precise and well-defined relations, such as *award received.* Moreover, they have been curated using manual and automated techniques, and hence, their truthfulness is easy to verify. Moreover, both variants of ChatGPT's probes perform significantly better than variants of Alpaca on correctness in both domains, with an out-performance of up to 36% in correctness@1. We also notice that, for both Alpaca and ChatGPT, their few-shot probes perform better than the zero-shot probes, with an improvement of 16% for Alpaca and 5% for ChatGPT. Finally, we find that many of the generated statements by humans and LLMs were actually statements with negative keywords but a positive meaning, such as *lebanon isn't devoid of historical sites*, with up to 14% of generated statements for the former and 11% for the latter. More samples are in Appendix C.

Salience of Truly Negative Statements. Results for salience are shown in Table 2. This metric is only computed over (previously annotated) correct statements. The best performances are shared between the *KG inferences* model and ChatGPT's few-shot variant. Though not performing comparably well overall, the *text extractions* model ranks first on salience of encyclopedic subjects @3 and 5. This is especially apparent for prominent entities, which are frequently queried using famous search engines. Again, ChatGPT's variants significantly outperforms Alpaca's on the notion of salience, with up to 23% improvement in salience@1, maintaining the same level of quality for both types of subjects. Sample results from all models are shown in Table 1 and Appendix E. An experiment on the quality of generated negatives over two popularity levels, namely prominent and long tail subjects, is in Appendix B.

Effect of k Value on LLM's Few-shot Probe. We examine the LLM using different numbers of samples, for in-context learning. We consider a subset of 5 entities (3 encyclopedic and 2 commonsense), and assess the performance of the few-shot ChatGPT using different values of k, with different salient:nonsalient ratios. Results are in Table 4. Adding a *small* but equal number of salient and nonsalient samples (3:3) improves the correctness by 8%, compared to only adding salient samples (3:0), however, at the expense of their salience, which drops by by 14%. Adding only nonsalient samples (0:3) compromises both metrics. Finally, adding a *larger* but equal number of salient and nonsalient samples (10:10) does not result in any improvements.

4. Take-home Lessons & Open Issues

In this paper, we perform a systematic evaluation of LLMs' ability to generate salient negative statements. We assess them against existing method and crowdsourced statements. We find that LLMs' few-shot probes show promising results in salience@1. Moreover, we find that ChatGPT outperforms Alpaca on this task, in both correctness and salience. One of the remaining limitations, however, is the ability of LLMs to recognize truly negative factual statements, as opposed to ambiguous, or seemingly negative statements with positive meaning. We hope that this study, as well as the following observations, give insights to future researchers on this topic.

Prompt Engineering. There is a wide consensus that LLMs are very powerful *when you ask them for information in the right manner.* In our task, we notice that the wording, especially of the zero-shot probe, changes the results dramatically. For instance, using the expressions *negative statements*, *negated statements*, and *negation statements* returns completely different responses. For instance, the probe with the word *negated* (alone without *salient factual*) returns obviously true statements with negative keywords added to them, e.g., "*stephen hawking was not a physicist*". The probe with the word *negative* does not return any results, but an apology from the AI about not being able to give *bad statements* about individuals. On this and other tasks, designing intuitive prompts and studying the ability of LLMs to understand them is the most important part of the process [17].

The Notion of Salient Negation. Assessing the truthfulness of statements is one thing, but assessing the salience of negatives is more challenging. Salience is a subjective metric. For instance, for a Basketball fan, the fact that *jordan* did not star in the film *space jam 2* (the first was built around him), is a big deal. For others, the salience is not obvious. In addition of the expertise of the reader, their nature is also important. In other words, are these negations generated for a human-reader, or to equip machines with better negative knowledge? For instance, what might not appear salient to a human, can be important to improve the reasoning skills of a chat bot. In this study, we assume that the reader is a human, who usually has a higher standard for what is interesting than a machine. Generally, designing experiments should take into consideration downstream applications and information about the end-user.

Maintenance. Ideally, models must always keep track of real-world changes which affect the truthfulness of statements, coverage of emerging entities, etc. This is relatively easy in the collaborative knowledge graphs, which are updated on a daily basis. For LLMs, the process of re-training is much more expensive. e.g., in May 2023, ChatGPT still generates the statement *brendan fraser has never won an oscar*, which is no longer true, due to his win in 2023 (the training of the model has been completed in September 2021).

References

[1] S. Razniewski, H. Arnaout, S. Ghosh, F. Suchanek, Completeness, recall, and negation in open-world knowledge bases: A survey, arXiv (2023).

[2] H. Arnaout, S. Razniewski, G. Weikum, J. Z. Pan, UnCommonSense: Informative negative knowledge about everyday concepts, in: CIKM, 2022.

[3] H. Arnaout, S. Razniewski, G. Weikum, Enriching knowledge bases with interesting negative statements, in: AKBC, 2020.

[4] T. Safavi, J. Zhu, D. Koutra, NegatER: Unsupervised Discovery of Negatives in Commonsense Knowledge Bases, in: EMNLP, 2021.

[5] J. Romero, S. Razniewski, K. Pal, J. Z. Pan, A. Sakhadeo, G. Weikum, Commonsense properties from query logs and question answering forums, in: CIKM, 2019.

[6] G. Karagiannis, I. Trummer, S. Jo, S. Khandelwal, X. Wang, C. Yu, Mining an "anti-knowledge base" from Wikipedia updates with applications to fact checking and beyond, PVLDB (2019).

[7] N. Lee, B. Z. Li, S. Wang, W.-t. Yih, H. Ma, M. Khabsa, Language models as fact checkers?, in: ACL, FEVER workshop, 2020.

[8] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: EMNLP, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI technical report (2019).

[11] N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, in: ACL, 2020.

[12] J. Chen, W. Shi, Z. Fu, S. Cheng, L. Li, Y. Xiao, Say what you mean! large language models speak too positively about negative commonsense knowledge, arXiv (2023).

[13] OpenAI, Introducing chatgpt, https://openai.com/blog/chatgpt, 2022.

[14] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledge base, CACM (2014).

[15] T. Nguyen, S. Razniewski, J. Romero, G. Weikum, Refined commonsense knowledge from large-scale web contents, TKDE (2022).

[16] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Alpaca: A strong, replicable instruction-following model, https://crfm.stanford.edu/2023/03/13/alpaca.html, 2023.

[17] J. Jang, S. Ye, M. Seo, Can large language models truly understand prompts? a case study with negated prompts, in: Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop, 2023.

A. Encyclopedic and Commonsense Subjects

We consider 50 subjects of different domains, namely commonsense and and of different popularity, namely prominent and long tail (see Table 5).

	Encyclopedic	Commonsense
Prominent	stephen hawking, michael jordan, lebanon, michelle obama, microsoft, china, amazon, albert einstein, the beatles, elon musk, angela merkel, taxi driver, taj mahal, white house, eat pray love, tom cruise, brendan fraser, the godfather, my cousin vinny, mercedes-benz group, gmc, linkedin	elephant, soup, lawyer, acne, mother, gorilla, pancake, newspaper, jaguar, avocado, garlic, chef, salad, rabbit, jogging, cufflink, strudel, librarian, armchair
Long tail	peri gilpin, caramel, ubisoft	tabbouleh, breadfruit, kitchenette, hockey stick, basketball court, coffee table

Table 5
Subjects considered in our experiments.

B. Prominent and Long Tail Subjects

We recompute the quality of negatives (@5) over two levels of subject-popularity, namely prominent and long tail. Figure 1 indicates a significant decrease in both salience and correctness for long tail subjects, for the text-based method; dropping to only 1% on salience. Using query logs as the corpus, users query prominent/trendy subjects much more frequently than long tail ones. We find the human-written statements for both popularity-levels comparable, with a slight advantage for prominent subjects. Similarly, the *KG inferences* model shows comparable results with a slight advantage of prominent subjects in correctness, and of long tail subjects in salience. Finally, we find an unexpected improvement, for all LLM variants, of long tail subjects over prominent ones, in both metrics. One interpretation could be the large amount of *noisy* web sources (main data source for training LLMs), about famous entities. For example, *tabbouleh* (long tail) is a specific instance of *salad* (prominent). While negatives about the former are more clear-cut, e.g., *tabbouleh isn't made with rice but bulgur*, negatives about the latter seem more unfocused, e.g., *salad isn't always a healthy choice.*

C. Negative Statement with Positive Meaning

As shown in Table 3, many of the LLM-generated and crowdsourced statements are in fact positive. Some of the recurring expressions which convey a positive meaning using negative keywords:

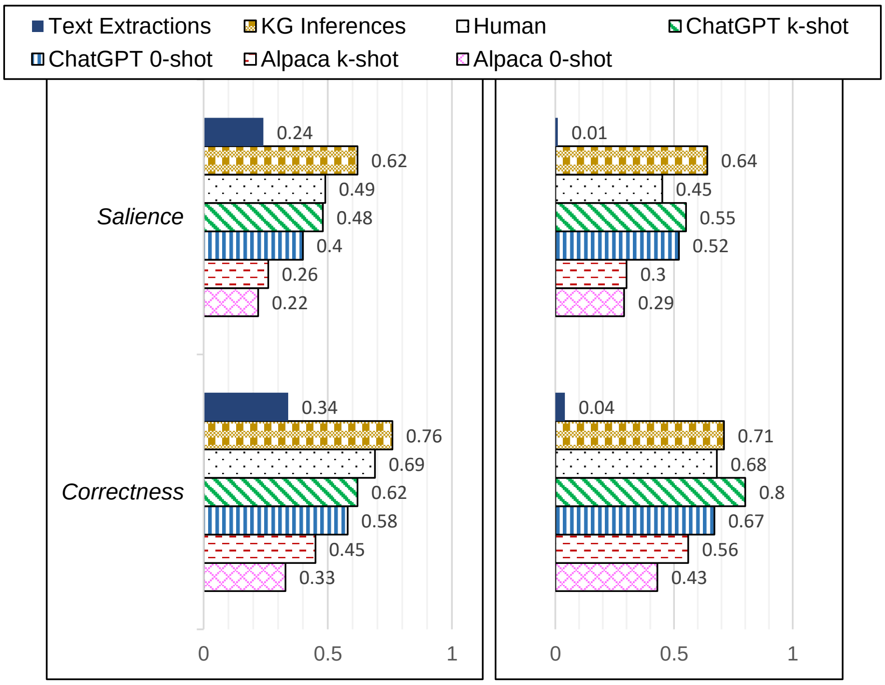

Figure 1: Prominent subjects (left-side), long tail subjects (right-side) (**best performance**, second best)

Expression: not exclusively (15 statements)
Amazon did not exclusively focus on selling its own products.

Expression: not without (2)
Example: *Strudel is not tasty without sugar.*

Expression: not just (9)
Example: *Acne is not just a teenage problem.*

Expression: not only (20)
Example: *Librarians do not only work in public libraries.*

Expression: not limited to (5)
Example: *Coffee tables are not limited to indoor use.*

Expression: not solely (7)
Example: *GMC does not solely operate in the United States.*

Expression: not all (10)
Example: *Not all librarians are women.*

D. k-shot In-context Learning Probe

In this probe k=6 (3:3); LLM=ChatGPT.

> *A salient factual negated statement about an entity means that the statement doesn't hold in reality. Moreover, the negated statement is either surprising, unexpected, or useful to the reader. For example:*
>
> `penguins can't fly.`
> `istanbul isn't the capital of turkey.`
> `tom cruise never won an oscar.`
>
> On the other hand, the following examples are factual negated statements that are not salient:
>
> `penguins can't run for presidency.`
> `istanbul isn't the capital of france.`
> `tom cruise never won the nba best player award.`
>
> *Given this definition and examples, write a list of* `3` *salient factual negated statement about* `microsoft`.
>
> **Answer:**
>
> 1. *is not primarily a dating platform.*
> 2. *does not charge users a fee to create an account.*
> 3. *does not allow users to post anonymous content.*

E. Sample Results

The following tables show the top results about *linkedin*, *chef*, and *angela merkel*, respectively:

Model	*Top Negative Statements (linkedin)*
Text	*isn't working* *isn't loading*
KG	*isn't headquartered in san francisco* *isn't a software company*
ChatGPT *0-shot*	*isn't designed for sharing personal content* *doesn't permit users to buy followers*
ChatGPT *k-shot*	*isn't used for online dating* *doesn't allow users to post pictures of their pets*
Alpaca *0-shot*	*doesn't have a user-friendly interface* *doesn't provide any value to its users*
Alpaca *k-shot*	*isn't a social media platform* *doesn't own the content posted on its platform*
Human	*doesn't have a billion members* *wasn't founded by mark zuckerberg*

Model	*Top Negative Statements (chef)*
Text	*doesn't wear hat* *doesn't eat their own food*
KG	*doesn't take orders* *doesn't bring drinks*
ChatGPT *0-shot*	*didn't use any garlic* *didn't win any cooking competitions*
ChatGPT *k-shot*	*doesn't just cook food* *not all have formal culinary training*
Alpaca *0-shot*	*don't need to have an understanding of nutrition* *don't need to have good knife skills*
Alpaca *k-shot*	*don't need to be certified* *don't usually work with raw ingredients*
Human	*doesn't wash the dishes* *doesn't always wear the chef's hat*

Model	*Top Negative Statements (angela merkel)*
Text	*didn't listen to donald trump* *doesn't deserve to be honoured by germany*
KG	*isn't on twitter* *isn't a lawyer*
ChatGPT *0-shot*	*isn't a native german speaker* *didn't originally pursue a career in politics*
ChatGPT *k-shot*	*has never been married* *is not a member of the SPD*
Alpaca *0-shot*	*isn't a member of the CDU* *isn't a scientist*
Alpaca *k-shot*	*isn't the first female chancellor of germany* *isn't from east germany*
Human	*didn't grow up in a wealthy family* *isn't a member of the SPD*