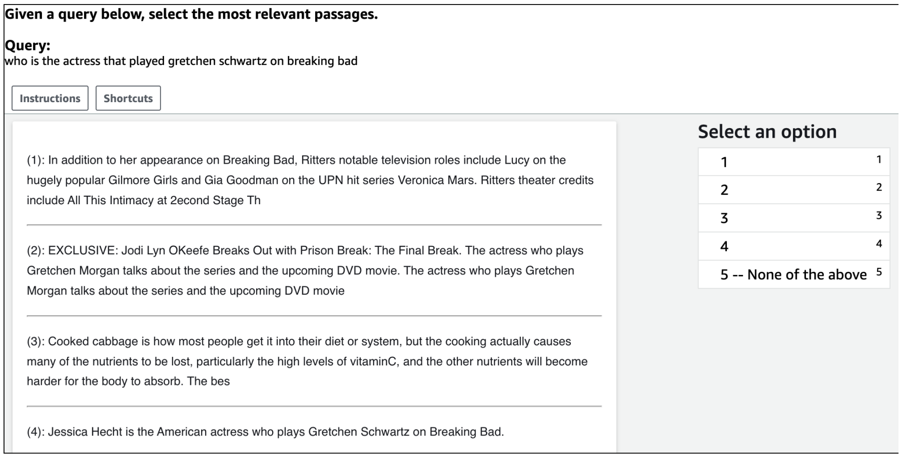
Active Learning with Crowd Sourcing Improves Information Retrieval: Appendix

A Details on Human Annotations

In this work, we used Amazon Mechanical Turk (MTurk) to collect crowd-sourced feedback. Fig. 5 shows our human intelligence task (HIT) design. In addition, we provided a short instruction along each task as shown in Fig. 6. Each task requires a human annotator to choose the most relevant passage from four possible candidates. Since those candidates are extracted as the top-4 answers from our model, there are no guarantees that the true answer is contained in those four candidates. Therefore, we allowed the human annotator to select "None of the above" if no relevant passages were present.


Given a query below, select the most relevant passages.

Query:
who is the actress that played gretchen schwartz on breaking bad

(1): In addition to her appearance on Breaking Bad, Ritters notable television roles include Lucy on the hugely popular Gilmore Girls and Gia Goodman on the UPN hit series Veronica Mars. Ritters theater credits include All This Intimacy at 2econd Stage Th

(2): EXCLUSIVE: Jodi Lyn OKeefe Breaks Out with Prison Break: The Final Break. The actress who plays Gretchen Morgan talks about the series and the upcoming DVD movie. The actress who plays Gretchen Morgan talks about the series and the upcoming DVD movie

(3): Cooked cabbage is how most people get it into their diet or system, but the cooking actually causes many of the nutrients to be lost, particularly the high levels of vitaminC, and the other nutrients will become harder for the body to absorb. The bes

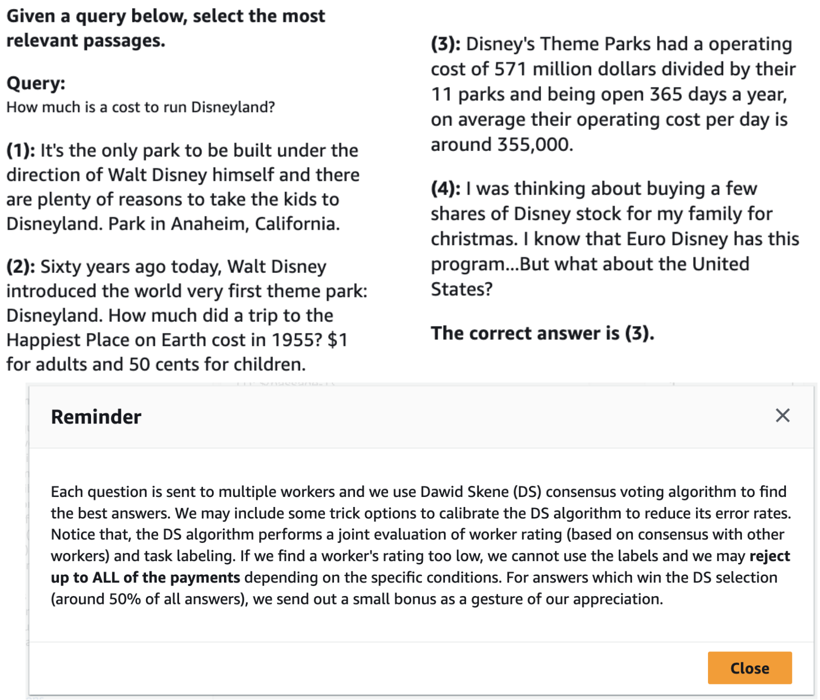
(4): Jessica Hecht is the American actress who plays Gretchen Schwartz on Breaking Bad.

Figure 5: Human intelligence task design on MTurk. Four candidates are displayed in random orders, each of which is truncated to at most 250 characters. A "none of the above" option is provided to reduce labeling noise.

An important aspect of HIT task is the design of instructions. Figure 6 shows an example of desired labels as well as additional reminders that we may disapprove payments as a mechanism to control for label qualities. We paid 0.45 USD for 3 annotators per task and used the following criteria for workers to ensure high quality: *Location: US; Minimal approval rate: 95%; Number of HITs approved: 1000.* One of our Turkers suggested to up the approval rate to 99% and increase the approval number to 10k while removing the "masters" badge for qualifications. We include this suggestion for future work.

A.1 Alternative Solutions

Before using MTurk, we started with Amazon SageMaker Ground Truth (GT) as our label collection system. Fig. 7 shows our interface design. In each task, a query and four relevant passages were presented to the human annotator, and a "None of the above" option was allowed. In addition, we used the Shapley Value method to extract keywords from each passage corresponding to the given query. Given a bi-encoder model, the keyword extraction was computed based on the cosine similarity between the given query and tokens from passages.



Given a query below, select the most relevant passages.

Query:
How much is a cost to run Disneyland?

(1): It's the only park to be built under the direction of Walt Disney himself and there are plenty of reasons to take the kids to Disneyland. Park in Anaheim, California.

(2): Sixty years ago today, Walt Disney introduced the world very first theme park: Disneyland. How much did a trip to the Happiest Place on Earth cost in 1955? $1 for adults and 50 cents for children.

(3): Disney's Theme Parks had a operating cost of 571 million dollars divided by their 11 parks and being open 365 days a year, on average their operating cost per day is around 355,000.

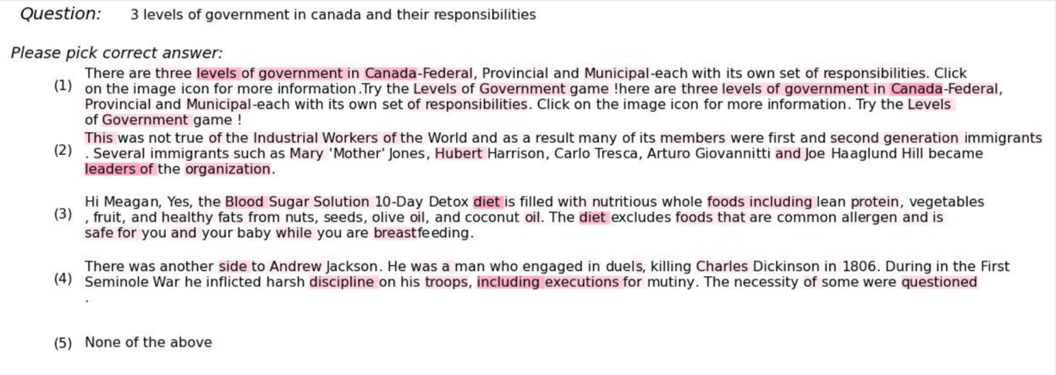
(4): I was thinking about buying a few shares of Disney stock for my family for christmas. I know that Euro Disney has this program...But what about the United States?

The correct answer is (3).

Reminder

Each question is sent to multiple workers and we use Dawid Skene (DS) consensus voting algorithm to find the best answers. We may include some trick options to calibrate the DS algorithm to reduce its error rates. Notice that, the DS algorithm performs a joint evaluation of worker rating (based on consensus with other workers) and task labeling. If we find a worker's rating too low, we cannot use the labels and we may **reject up to ALL of the payments** depending on the specific conditions. For answers which win the DS selection (around 50% of all answers), we send out a small bonus as a gesture of our appreciation.

Figure 6: Instruction and reminder in Human intelligence task design on MTurk.



Question: 3 levels of government in canada and their responsibilities

Please pick correct answer:

(1) There are three levels of government in Canada-Federal, Provincial and Municipal-each with its own set of responsibilities. Click on the image icon for more information.Try the Levels of Government game !here are three levels of government in Canada-Federal, Provincial and Municipal-each with its own set of responsibilities. Click on the image icon for more information. Try the Levels of Government game !

(2) This was not true of the Industrial Workers of the World and as a result many of its members were first and second generation immigrants . Several immigrants such as Mary 'Mother' Jones, Hubert Harrison, Carlo Tresca, Arturo Giovannitti and Joe Haaglund Hill became leaders of the organization.

(3) Hi Meagan, Yes, the Blood Sugar Solution 10-Day Detox diet is filled with nutritious whole foods including lean protein, vegetables , fruit, and healthy fats from nuts, seeds, olive oil, and coconut oil. The diet excludes foods that are common allergen and is safe for you and your baby while you are breastfeeding.

(4) There was another side to Andrew Jackson. He was a man who engaged in duels, killing Charles Dickinson in 1806. During in the First Seminole War he inflicted harsh discipline on his troops, including executions for mutiny. The necessity of some were questioned .

(5) None of the above

Figure 7: Human intelligence task design on Amazon Ground Truth.

Pros: GT offers a programmatic solution to the labeling problem, where we provide the tasks and obtain the results. It was easy to set up, allowed full automation, and was our gateway to crowd sourcing. From our experience, we obtained reasonably good results when the tasks could be completed in a few seconds and the true answers were obvious. Examples include tasks containing visual cues, such as image similarity, or answers containing only a few terms. While some human labels were notably noisy in these tasks, we were able to obtain results that met our expectations in the aggregate. As a result, we could use GT to mock online A/B tests for simple problems.

Cons: GT does not offer direct connections to the labelers or consider payment rejections, but instead approves all labels automatically. Additionally, as can be seen from the examples in Figures 5 and 7, search ranking tasks can not easily be completed in a few seconds, but instead cost an average of 30 seconds to complete by our first-hand experience. As a result, labelers for GT likely have more incentives to game the system by providing random answers. We measured 23% to 25% accuracy from the human responses, based on ground-truth labels when the correct answers were included in the candidate sets. The accuracy is near random one-in-four selections, invalidating these labels.

To address these limitations, we chose MTurk as our preferred environment for data collection. By choosing labelers with good approval histories and using a default option to delay the payments by 72 hours, we already observed significant improvements in label quality, presumably due to psychological factors. We further implemented latent-variable rating systems (Section 2.4) and added clear instructions (Figure 6), which allowed us to directly measure and provide guarantees for label qualities.

Alternatively, Amazon SageMaker Ground Truth offers a Plus version where the labelers may be provided through contracts for more transparency and confidentiality. We did not explore this option because contract negotiations, especially involving third-party vendors, may require additional work. Our primary goal was to keep the approach open and reproducible. Our results showed that Amazon Mechanical Turk could yield good labels with our implemented filtering and rating systems.

B Discussions of Detailed Results

Besides our main experimental results, we here discuss two useful visualization tools regarding model improvements and labeler qualities. We developed these tools when we were unsure of the final performance outcomes and they greatly helped us analyze our approach.

B.1 Knowledge Transfer Visualization

As part of our generalization claim, we aim to explain which label (q_0, d_0) enables the fine-tuned model to retrieve the correct answer for an unseen query (q_*, d_*). This is done by finding positive changes to a utility function for the target pattern, $u(q_*, d_*, \boldsymbol{\theta}_*)$. Usually, explainability requires one to consider the complex relations of all training examples. Here, we simplify the approach with only marginal changes from a one-step gradient of model parameters:

$$\boldsymbol{\theta}_* = \boldsymbol{\theta}_0 + \eta \partial_{\boldsymbol{\theta}} (u(q_0, d_0, \boldsymbol{\theta}))|_{\boldsymbol{\theta}_0},$$

where η is the step size, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_*$ represent the model parameters before and after update respectively. By inserting the update to the prediction of the target pattern, we see the relative change in the utility score,

$$\begin{aligned} & u(q_*, d_*, \boldsymbol{\theta}_*) - u(q_*, d_*, \boldsymbol{\theta}_0) \\ & = u(q_*, d_*, \boldsymbol{\theta}_0 + \eta \partial_{\boldsymbol{\theta}} (u(q_0, d_0, \boldsymbol{\theta}))|_{\boldsymbol{\theta}_0}) - u(q_*, d_*, \boldsymbol{\theta}_0), \\ & \approx \eta \partial_{\boldsymbol{\theta}} (u(q_*, d_*, \boldsymbol{\theta})|_{\boldsymbol{\theta}_0})^T \partial_{\boldsymbol{\theta}} (u(q_0, d_0, \boldsymbol{\theta})|_{\boldsymbol{\theta}_0}). \end{aligned} \tag{10}$$

Therefore, if (q_0, d_0) is similar to (q_*, d_*), we may observe high dot-product scores of their gradients, which indicates a positive gain to the utility function in the approximation.

In Figure 8, we show a query “the resource bank routing number”, which has two components: “resource bank” and “routing number”. The zero-shot model failed to recognize the correct meaning of these objectives. It favored answers that contain exact matches of the entity instead of answers that explain the query. In this case, it incorrectly identified “resource” as assets. When the zero-shot model was fine-tuned on a training dataset that does not contain this specific query, it returned the correct answer. By performing the gradient similarity search in Eq. (10) on all training samples, we identified two query-answer pairs that contributed the most to the utility score of the target query-answer pair. As shown in Figure 8, both relevant queries have the same structure as the target query, “The bank routing number”, which provides semantic information to the zero-shot model to generalize the knowledge on “Resource bank routing number”.

B.2 Dawid-Skene Voting and Rating

Figure 9 shows the DS evaluation score and the labeler’s chance to choose the corrupted answers, which can be up to 0.2 if the labeler chooses answers completely randomly. We see a high correlation between the two methods, showing that the DS score is a reliable way to assess labeler qualities. We set a threshold of $\gamma = 0.15$ as our payment rejection criterion.

The rejection rate in Figure 9 was 13.2% by the number of labels. Our highest rejection rate was 27% in our initial iteration, yet we attribute that to our lack of reminders in the instructions and we reconciled the labelers who reach out to us. Our lowest rejection rate was 0%, yet this can be a rare

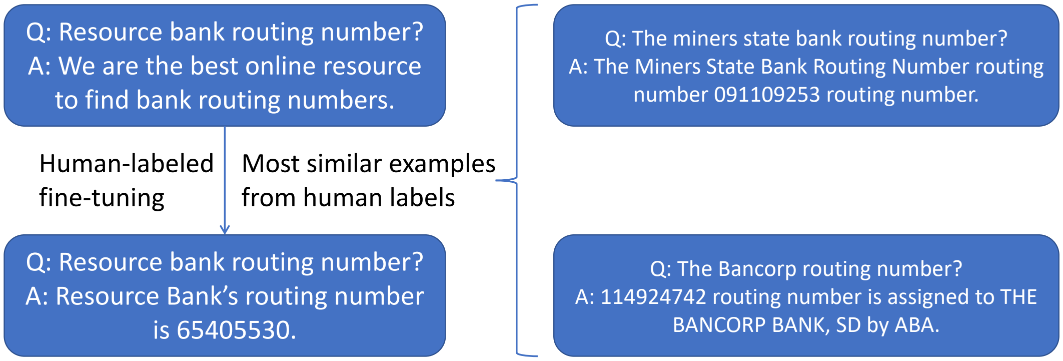

Q: Resource bank routing number?
A: We are the best online resource to find bank routing numbers.
Human-labeled fine-tuning
Most similar examples from human labels
Q: Resource bank routing number?
A: Resource Bank's routing number is 65405530.
Q: The miners state bank routing number?
A: The Miners State Bank Routing Number routing number 091109253 routing number.
Q: The Bancorp routing number?
A: 114924742 routing number is assigned to THE BANCORP BANK, SD by ABA.

Figure 8: An example for knowledge transfer visualization

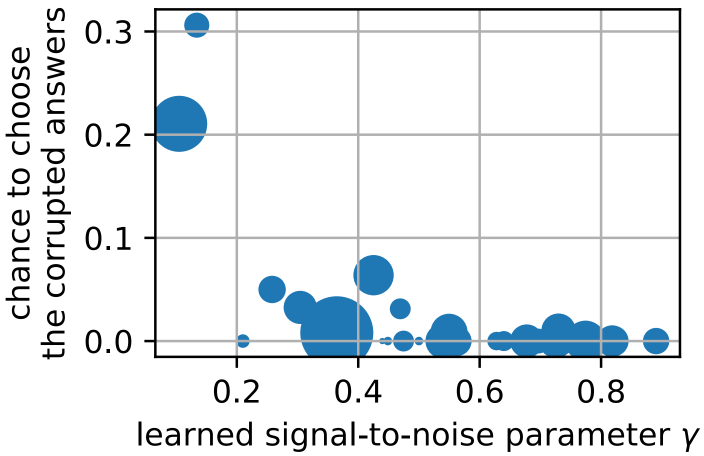


Figure 9: Dawid-Skene calibration against ground-truth labeler noise, collected from Iteration 2 in NQ experiments. Each dot represents a labeler and its size shows the number of labels they provided. A high correlation can be seen between the two methods.

phenomenon. The rejection rate is random in nature and it can highly depend on the time of day and other external factors. The rejected answers were re-released to gather clearer signals, though we did not find significant changes in fine-tuning performance. Besides labeler attribution, one side benefit from DS algorithm is the ability to produce single answers when majority voting produces ties for over 50% of our tasks.

C Examples from Human Evaluation Tasks

Tables C1 to C4 show examples from our final human evaluation tasks to provide more insights into the challenges of active learning in the real world. We include cases for each winning method, where our proposed CCR with human labels had the highest winning rates, shown in Figure 4 in Section 3.4. For cases where human selected "none of the above" options, we also included answers from the original datasets. Notice that CCR with human labels never had access to these labels and CCR with oracle labels only had access if these answer passages were retrieved in one of the active learning steps. Therefore the retrieval of the optimal answers is fundamentally challenging. Our work shows that we can use pretrained models with a modest budget (equivalent to 2% of what was used to collect the original labeled datasets) to demonstrate effectiveness in active learning with human feedback for cold-start search problems.

D Additional References

Crowdsourcing in Active Learning The domain of active learning utilizing crowdsourced contributors has a long history [35]. In Pfeiffer et al. [42], the query is a pair of images with dots and the crowdsourced workers are asked to identify the image with the most dots. In Chen et al. [11], the query is a pair of documents, and the crowdsourced workers are asked to rank them by reading difficulty. In both works, the queries were adaptively selected. Our work differentiates itself primarily in terms of the domain, the complexity of the neural network models employed, and the difficulty of the tasks, which are typically handled by trained labelers.

Human Feedback on Language Modeling Pre-trained language models such as GPT-3 [44] and Facebook's XLM and XLM-R [13] have significantly advanced the field of natural language understanding. However, it has been observed that the behavior of these large language models does not always align precisely with user intentions [40]. This discrepancy can be attributed to both biased training objectives and data. For instance, a contrastive loss may lead to popularity bias, where frequently occurring answers deviate from users' actual preferences [63]. Furthermore, training data in the language domain, often constructed with assistance from other search engines [39], can introduce undesired bias into the training process. Studies have shown that aligning the output of a trained language model with human behavior can be achieved via fine-tuning with human feedback [40]. Notably, recent work on ChatGPT [40] and related topics have underscored the considerable potential of human-in-the-loop machine learning. One way to appreciate the success of ChatGPT is to understand it as a pre-fine-tuned model in common domains of interest. in the specific domains related to search and recommendation, is still relatively unclear how to fine-tune the models based on human feedback. For example, Gao et al. [20, 21]; Zhang et al. [64] documented various attempts at using pre-trained language models for direct movie recommendation. While these efforts provide a solid starting point, we posit that incorporating human feedback can further enhance domain adaptivity, a topic we aim to explore in this paper.

Query	what are the 6 parts of the brain
Human	There are 8 parts of the brain. Heres a list of what they are and what they do. The frontal lobe is one of the 4 major divisions of the cerebral cortex. This part of the brain regulates decision making, problem solving, control of purposeful behavior
Oracle	What are the parts of the brain? The main parts of the brain are the frontal lobe, temporal lobe, occipital lobe, parietal lobe, hypothalamus, cerebrum, brain stem ,and cerebellum. In addition, the brain contains the corpus callosum, pituitary gla
BM25	Here are some examples of functions that the brain controls: The brain is like a busy city. Each part has different functions and is made up of different types of cells. To work, different parts of the brain need to send messages to each other, and t
Zero-Shot	What Are the Parts of the Brain? Every second of every day the brain is collecting and sending out signals from and to the parts of your body. It keeps everything working even when we are sleeping at night. Here you can take a quick tour of this amaz

Query	synonym for the word evaluate
Human	Synonyms for evaluate in the sense of this definition. (evaluate is a kind of ...) use or exercise the mind or ones power of reason in order to make inferences, decisions, or arrive at a solution or judgments.
Oracle	Definition 1: evaluate or estimate the nature, quality, ability, extent, or significance of [verb of cognition] Samples where evaluate or its synonyms are used according to this definition. Synonyms for evaluate in the sense of this definition. (eva
BM25	Definitions and Synonyms of evaluate Another word for evaluate What is evaluate? Definition 1: evaluate or estimate the nature, quality, ability, extent, or significance of [verb of cognition] Samples where evaluate or its synonyms are used accord
Zero-Shot	Here are all the possible meanings and translations of the word evaluate. Princetons WordNet(5.00 1 vote)Rate this definition: measure, evaluate, valuate, assess, appraise, value(verb) evaluate or estimate the nature, quality, ability, extent, or si

Query	what is a straddle
Human	Straddle(verb) to place one leg on one side and the other on the other side of; to stand or sit astride of; as, to straddle a fence or a horse. Straddle(noun) the act of standing, sitting, or walking, with the feet far apart. Straddle(noun)
Oracle	A straddle is any set of offsetting positions on personal property. One example, is a put and call option on the same number of shares of a particular security, with the same exercise price and expiration date.
BM25	straddle my rail; Straddle Piss; straddle puss; straddler; Straddle Racking; Straddlers Twat; Straddleships; straddle shit; Straddle some cows; straddle tale; Straddle the Bench; straddle the fence; straddle the gauze; Straddlewhipped; straddle worth
Zero-Shot	Define straddle. straddle synonyms, straddle pronunciation, straddle translation, English dictionary definition of straddle. v. straddled , straddling , straddles v. tr. 1. a. To stand or sit with a leg on each side of; bestride: straddle a horse. b.

Query	stonewalling definition
Human	Definition: When a spouse is stonewalling in communication in a marriage relationship, he or she is usually.1 using delaying or stalling tactics, or. 2 refusing to answer questions, or. 3 doing whatever can be done to hinder or obstruct a discussi
Oracle	Stonewalling is a refusal to communicate or cooperate. Such behaviour occurs in situations such as marriage guidance counseling, diplomatic negotiations, politics and legal cases. Body language may indicate and reinforce this by avoiding contact and
BM25	stonewall riot. Stonewall Riot definition. A disturbance that grew out of a police raid on the Stonewall Inn, a popular hangout for gays in Manhattans Greenwich Village in 1969. Such raids long had been routine, but this one provoked a riot as the cr
Zero-Shot	Examples of stonewall in a Sentence. 1 They stonewalled until they could come up with a response. 2 They were just stonewalling for time. 3 Theyre trying to stonewall the media. 4 Were trying to get the information, but were being stonewalled.

Table C1: Examples of final human evaluation tasks on MSMARCO dataset (Part 1 of 2). The answers retrieved from the highlighted methods were chosen as the correct labels.

Query	how many calories do i burn to lose weight
Human	How many calories a day do you need to burn to lose weight? A: Calorie Secrets states that 1 pound of fat is equal to 3,500 calories. Therefore, in order to lose 1 pound of body fat you must burn 3,500 more calories th... Full Answer
Oracle	How many calories do I need to burn to lose just one pound? 3500 calories: You have to cut down 3500 calories in your diet to loose one pound if you cut down 500 calories per day, you will loose 1 pound in on week, and approx 4 lbs in a month. ...Rea
BM25	A 3500 Calorie Deficit Approximately 1 Pound of Weight Loss. If you are trying to lose weight through calorie counting, you will need to know how many calories it will take to lose a pound of weight. Creating a calorie deficit of about 3500 calories
Zero-Shot	So the amount of energy you exert in doing an activity is measured by the calories burn rate. How to burn calories? Thats easy, just be alive! Your body is constantly burning calories to keep your body functioning. To burn more calories, do more acti

Query	how did the uluru rock form
Human	UluruAyers Rock is a rock in the Northern Territory in Australia. It is made of red sandstone. It is not the largest rock in the world, being second to Australias Mt Augustu s, which is almost twice the size.yers Rock, now known by its original name
Oracle	Uluru, formerly known as Ayers Rock, is a large rock located in the Northern Territory. Ayers Rock was named after the 19th century Premier of South Australia, Sir Henry Ayers.It is located in UluruKata Tjuta National Park, 350 kilometres southwest o
BM25	When did Uluru become a national park? In 1950 Ayers Rock, today known as Uluru, was declared a national park. In 1958 both Ayers Rock and Mt Olga (Kata Tjuta) were excised from an Aboriginal reserve to form the Ayers Rock Mt Olga National Park.
Zero-Shot	Uluru is easily the most iconic natural landform in Australia, and its formation was equally special. The creation of Uluru and Kata Tjuta as both were formed at the same time began over 500 million years ago.At this time the big crustal blocks tha

Query	does quotation mark go before or after period
Human	1 When the whole sentence except for the section enclosed in quotation marks is a question or exclamation, the question or exclamation mark goes outside the quotation mark. 2 When only the unit in quotation marks is a question or exclamation, the ma
Oracle	Proper placement of the period with quotation marks. If a sentence ends with quoted material, the period is placed inside the closing quotation mark, even if the period is not part of the original quotation. Note, however, that if the quoted material
BM25	after the quotation marks because if put before the quotation mark, that makes the quote seem like if it continues after what you wrote even if the quote has ended. period mar ks go before the quotation mark because that is ending a sentence... peri
Zero-Shot	Do period marks come after parenthesis or before? In a sentence like this, does the period mark come before or after parenthesis? I walked to the door (but I didnt see it was closed). or I walked to the door (but I didnt see it was closed.) What abou
None of the above (Ground Truth 1)	If the quote is the complete sentence in itself, then the period goes inside the quotation mark. If the quote is part of a larger sentence, then the period goes after the quotation mark. Here is an example: I spoke to her and she told me I don't like

Table C2: Examples of the final human evaluation tasks on MSMARCO dataset (Part 2 of 2). The answers retrieved from the highlighted methods were chosen as the correct labels. For completeness, we also include examples where human selected "None of the above" answer, where we reveal the ground-truth labels in the original dataset. See Appendix C for further discussions.

Query	when does the new season of law and order svu come on
Human	Law & Order: Special Victims Unit (season 4): Filming for Season 4 began while Season 3 was still airing as evidenced by reports that Sharon Lawrence would appear on SVU in time for May sweeps.[1][2]
Oracle	Law & Order: Special Victims Unit (season 19): Michael Chernuchin, who had previously worked on Law & Order, Law & Order: Criminal Intent, and Chicago Justice took over from Rick Eid as showrunner. This is also the first season since season twelve in
BM25	Law & Order: Special Victims Unit: Executive producer Neal Baer left Law & Order: SVU as showrunner at the end of season twelve, after eleven years (seasons 212) on the show, in order to sign a threeyear deal with CBS Studios.[11] Baer was replaced b
Zero-Shot	Law & Order: Special Victims Unit: In 2016, a New York Times study of the 50 TV shows with the most Facebook Likes found that SVUs popularity was atypical: generally slightly more popular in rural areas and the Black Belt, but largely restricted to t

Query	who won the peloponnesian war and how did they win
Human	Peloponnesian War: Sparta and its allies, with the exception of Corinth, were almost exclusively landbased powers, able to summon large land armies which were very nearly unbeatable (thanks to the legendary Spartan forces). The Athenian Empire, altho
Oracle	History of the Peloponnesian War: The History of the Peloponnesian War (Greek: , Histories) is a historical account of the Peloponnesian War (431404 BC), which was fought between the Peloponnesian League (led by Sparta) and the Delian League (led by
BM25	Melos and the Peloponnesian War: [4] When looking to find examples of realism, there is a definite bias that comes into play. This is one that arises from a desire to prove realism is an always evident paradigm that can explain past and future occurr
Zero-Shot	History of the Peloponnesian War: For the most part, the History does not discuss topics such as the art and architecture of Greece.

Query	when did the three little pigs come out
Human	The True Story of the 3 Little Pigs!: The True Story of the 3 Little Pigs! is a childrens book by Jon Scieszka and Lane Smith. Released in a number of editions since its first release by Harper & Row Publishers in 1989, and republished the name of Vi
Oracle	The Three Little Pigs: The Three Little Pigs was included in The Nursery Rhymes of England (London and New York, c.1886), by James HalliwellPhillipps.[1] The story in its arguably bestknown form appeared in English Fairy Tales by Joseph Jacobs, first
BM25	The Three Little Pigs: The third little pig builds a house of bricks. The wolf fails to blow down the house. He then attempts to trick the pig out of the house by asking to meet him at various places, but he is outwitted each time. Finally, the wolf
Zero-Shot	The True Story of the 3 Little Pigs!: This is the story of the 3 little pigs from the perspective of Alexander T. Wolf. The wolf is trying to set the story straight of how he came to be big and bad. At the beginning of the book, he is cooking a cake

Query	when does the movie the star come out
Human	The Star (2017 film): The first trailer was released on July 26, 2017.[20] On November 16, 2017, the official video for the song The Star, performed by Mariah Carey, was made available on her YouTube channel.[21]
Oracle	The Star (2017 film): In July 2016, the release date was set for November 10, 2017,[18] but it was later pushed back to November 17, 2017.[19]
BM25	Monsters University: Leonard Maltin of IndieWire praised the animation and art direction, but wrote that he wished the movie was funnier and wasnt so plotheavy and that Pixar has raised the bar for animated features so high that when they turn out a
Zero-Shot	Movie star: Movie stars in other regions too have their own star value. For instance, in Asian film industries, many movies often run on the weight of the stars crowd pulling power more than any other intrinsic aspect of film making.

Table C3: Examples of final human evaluation tasks on Natural Questions dataset (Part 1 of 2). The answers retrieved from the highlighted methods were chosen as the correct labels.

Query	who is the longest serving manager in manchester united history
Human	Alex Ferguson: Ferguson was appointed manager of Manchester United in November 1986. During his 26 years with Manchester United he won 38 trophies, including 13 Premier League titles, five FA Cups, and two UEFA Champions League titles.[9] He was knig
Oracle	Premier League: The leagues longestserving manager was Alex Ferguson, who was in charge of Manchester United from November 1986 until his retirement at the end of the 201213 season, meaning that he was manager for all of the first 21 seasons of the P
BM25	201213 Manchester United F.C. season: On 8 May 2013, Uniteds long time manager, Sir Alex Ferguson announced that he would retire from his position as manager of Manchester United after 26 and a half years in charge, making him the longestserving mana
Zero-Shot	Ryan Giggs: The son of rugby union, and Wales international rugby league footballer Danny Wilson, Giggs was born in Cardiff but moved to Manchester at the age of six when his father joined Swinton RLFC. Predominantly a left winger, he began his caree

Query	when does the new season of lost in space come out
Human	Lost in Space (2018 TV series): In October 2014, it was announced that Legendary Television was developing a new reboot of Lost in Space and had hired Dracula Untold screenwriting duo Matt Sazama and Burk Sharpless to pen the new series.[12] In Novem
Oracle	Lost in Space (2018 TV series): Lost in Space is an American science fiction television series based on a reimagining of the 1965 series of the same name (itself a reimagining of the 1812 novel The Swiss Family Robinson), following the adventures of
BM25	Lost in Space: In early 1968, while the final thirdseason episode Junkyard in Space was in production, the cast and crew were informally made to believe the series would return for a fourth season. Allen had ordered new scripts for the coming season.
Zero-Shot	Lost in Space (2018 TV series): Toby Stephens, speaking about the distinction between the original series and the new show:

	who plays rachel on jessie punch dumped love
Human	Julia Garner: Garner has also acted in another Netflix series, Maniac, as Ellie. Ellie is the sister of main character Annie, played by Emma Stone.[8]
Oracle	Jennifer Veal: Jennifer Anne Veal (born 7 September 1991) is a British actress and comedian from Coventry, England. She is best known for her work on YouTube formerly alongside Lucas Cruikshank, as well as her role as Agatha on the American televisi
BM25	Adam Sandler: In 2013, he guest starred in the Disney Channel Original Series Jessie as himself. He and Cameron Boyce previously worked together in Grown Ups and Grown Ups 2. The episode is titled Punched Dumped Love. Sandler costarred in the drama f
Zero-Shot	List of Saved by the Bell characters: From sophomore year until the end of the series, Jessie dates athlete A.C. Slater in an opposites attract relationship, which causes friction between the both of them. Slaters pet name for Jessie is Mama. Jessie
None of the above (Ground Truth 1)	List of Jessie episodes: Guest stars: Lenny Jacobson as Ted the Delivery Guy, Isabella Palmieri as Rachel Kapowski, Jackson Odell as Gale

Table C4: Examples of the final human evaluation tasks on Natural Questions dataset (Part 2 of 2). The answers retrieved from the highlighted methods were chosen as the correct labels. For completeness, we also include examples where human selected “None of the above” answer, where we reveal the ground-truth labels in the original dataset. See Appendix C for further discussions.