Could AI Leapfrog the Web? Evidence from Teachers in Sierra Leone

Authors: Daniel Björkegren[1]*, Jun Ho Choi[1], Divya Panchaksharappa Budihal[1], Dominic Sobhani[2], Oliver Garrod[3], Paul Atherton[3]

Affiliations:

[1]School of International and Public Affairs, Columbia University; New York, United States of America.

[2]Columbia University; New York, United States of America.

[3]Fab Inc.; London, United Kingdom.

*Corresponding author. Email: dan@bjorkegren.com

Abstract: Access to digital information is a driver of economic development. But although 85% of Africa's population is covered by mobile broadband signal, only 37% use the internet, and those who do seldom use the web. We investigate whether AI can bridge this gap by analyzing how 469 teachers use an AI chatbot in Sierra Leone. The chatbot, accessible via a common messaging app, is compared against traditional web search. Teachers use AI more frequently than web search for teaching assistance. Data cost is the most frequently cited reason for low internet usage across Africa. The average web search result consumes 3,107 times more data than an AI response, making AI 87% less expensive. Additionally, only 2% of results for corresponding web searches contain content from Sierra Leone. In blind evaluations, an independent sample of teachers rate AI responses as more relevant, helpful, and correct than web search results. These findings suggest that AI-driven solutions can cost-effectively bridge information gaps in low-connectivity regions.

Main Text:

The internet has come up short in sub-Saharan Africa. After a policy focus on building network infrastructure, mobile phone penetration is high (92 phones per 100 people), and 85% of the population has access to mobile internet signal (*1*). Yet only 37% of the population uses the internet (*2*). Even more striking is *how* this minority uses the internet: while 76% of internet users across five African countries used social networking daily in 2022, only 29% used web search daily, as shown in Figure 1 Panel A. Connectivity can improve economic outcomes in low-income countries (*3*, *4*), but this lack of adoption suggests that the full potential of the information revolution has yet to be realized.

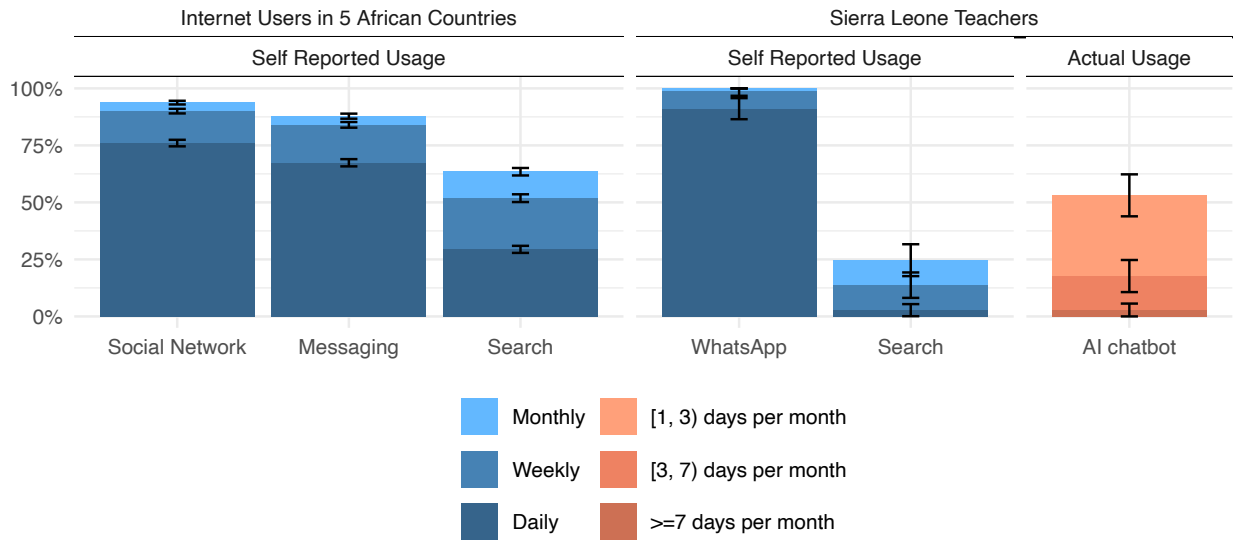


Fig. 1. Use of internet services in sub-Saharan Africa. The left panel shows self-reported usage among individuals who used the internet in the past three months, from a representative survey of households in Ethiopia, Kenya, Nigeria, South Africa, and Uganda in 2022 (*5*). The right panel shows usage in the sample of surveyed teachers in Sierra Leone. This includes endline self-reported usage of WhatsApp for texting, and web search for support in teaching activities. The final bar shows actual usage of the AI chatbot, in terms of average number of days with a query submitted per 30 days (month). Error bars indicate the 95% confidence intervals for that frequency or higher. See Figures S1 and S2 for other measures of usage.

We consider two large barriers that inhibit use of the web. First, web content rarely serves local needs. An audit study found that in sub-Saharan Africa fewer than 30% of a sample of web search results contained local content, compared to over 90% in North America (*6*).[1] Foreign content is typically formatted for viewing on large computer screens, connected via broadband—a poor fit for the mobile devices and unstable connections through which most Africans access the web. While mobile phones are common, just 17% of adults report owning a computer on average across 17 sub-Saharan African countries (*7*). Second, cost is a significant barrier. 77% of

[1] In that paper, search results were gathered for queries with the names of a set of worldwide capital cities.

internet users report that an issue limits their use of the internet; the most cited issue is data cost, reported by 66% of users who are limited. Network performance or availability is reported as the primary issue by 11% of users who are limited (*5*).[2] Under pay-as-you-go plans that are common, users must purchase each megabyte or gigabyte of data.

We consider how these barriers affect a group of knowledge workers who require reliable access to information. We consider teachers in Sierra Leone, a group that regularly uses messaging platforms like WhatsApp for communication but struggles to leverage the web. This study considers two questions. First, could AI make information more accessible to this population? Second, what does the use of AI reveal about information needs that are not currently being met?

We study the use of a WhatsApp-based AI chatbot for teachers in Sierra Leone, which allows teachers to submit queries to an AI agent as they would to a friend or family member. As shown in Fig. 1 Panel B, teachers given access use AI for teaching activities more than web search. Usage is sustained—for the earliest cohort, over 17 months. Basic descriptives on early usage of this system were described in a previous non-archival workshop paper (*8*). This paper considers whether AI usage can 'leapfrog' over web usage, filling information gaps that have not been filled by the web.

We use chatbot queries as a window into users' information needs, similarly to how previous studies have used web search queries, which are seldom used in this context (*9*). We submit these chatbot queries to the local version of Google search and compare the results to better understand the relative utility of web versus AI for a population of knowledge workers in a low-income country.

Context

This study focuses on the education sector within Sierra Leone. The country had a GDP per capita of $509.48 in 2021; by that measure, the average American is 137 times richer.

Teachers are knowledge workers who regularly must develop content (lesson plans, assignments, and assessments), and navigate complex interpersonal scenarios (with students, parents, and colleagues). Sierra Leonean teachers are stretched thin: there are 42 students per primary teacher, and they have few resources for professional development. The median teacher in our survey had only completed 2 continuous professional development trainings within the past 3 years. Only 19% of surveyed teachers hold bachelor's degrees.

Many teachers work in remote areas that have limited infrastructure. In 2020, only 34% of secondary schools had electricity, and just 8% had internet access. These numbers are even lower among primary schools (6% and 1%). But access to 2G (voice and SMS) is more widespread: 86% of schools nationally were within range of mobile phone coverage (*10*), and some teachers may have 3G signal in their homes.

As shown in the right panel of Fig. 1, in a survey of teachers, 91% use WhatsApp daily. They are not strangers to technology. However, they infrequently search. After being provided with an AI WhatsApp chatbot, only 3% report using web search daily to help with teaching, and only 25% report using search at least monthly. In contrast, 53% use the AI chatbot at least one of every 30 days on average over our data.

[2] In some countries, mobile phone operators offer data plans for social media, or that allow access to a basic text based version of Facebook and WhatsApp for free with a bundle.

AI Chatbot

Design

From the teacher's perspective, the AI chatbot (called 'theTeacher.AI') appears as a contact in WhatsApp, and they may chat with it just as with any other contact. The design of the system is similar to that of a chatbot for Kenyan entrepreneurs developed and analyzed in (*11*), but tailored for education.

On the backend, the chatbot relays the conversation, plus a system message, to the OpenAI API, and returns the result back to the user.[3] The system message tailors the system to act like a helpful education assistant, to follow best practice pedagogy (guided by the Education Endowment Fund's reports on education evidence (*12*)), to follow safeguards, and to adapt responses to a teacher who has few classroom resources. The chatbot answers only queries related to education due to budget and for quality control. It was accessible in English, the official language of Sierra Leone; 95% of surveyed teachers in our sample report that English is a main language spoken at home. However, the system message tailors responses to use short, simple language that would be readable to someone without English as a first language.

The core chatbot code is available in an open-source repository and the full system message is given in Appendix B1, so that the system can be replicated by others.[4]

Recruitment and Onboarding

Classroom teachers and principals were recruited in several cohorts from government schools and a non-profit school network (EducAid Sierra Leone). Users were given onboarding training which ranged from one to several hours depending on the cohort. Training including an overview of the system and practice asking questions. As part of onboarding, users were asked for informed consent for their anonymized data to be analyzed for research. We consider only teachers and principals using the system from Sierra Leonean phone numbers. We omit queries from each user's first day, to omit any practice questions submitted during the training session, and include only users who successfully submitted at least one query after the training day. This yields 469 users, of which 86% are classroom teachers and 14% are head teachers (principals). For simplicity, we refer to users of both types as teachers.

One cohort of teachers in the non-profit schools, which we call the survey sample (N=122), was surveyed three times by EducAid. They were surveyed during onboarding ('baseline', in winter 2024), 2-3 months after onboarding ('midline'), and 3-4 months after onboarding ('endline').

Usage

Figure 1 Panel B shows the leapfrogging pattern in the survey sample. For teaching activities, more teachers use AI each month than use web search. The figure reports AI usage from chatbot logs, and self reported search and WhatsApp usage. In Figure S2 we report alternate measures. Self-reported AI usage is even higher. More teachers report consulting the AI chatbot for information about teaching (77.4%) than the internet (37.0%); the proportion reporting consulting the chatbot is not statistically significantly different from the proportion consulting

[3] The usage captured during this study was routed to GPT 3.5 Turbo, a state of the art model at the time of the first deployment; the system has since been upgraded to GPT 4o mini.
[4] See https://github.com/AI-for-Education/fabdata-llm

books (68.5%). High usage of an AI chatbot is not unique to our intervention; it is also observed in a study with Kenyan entrepreneurs (*11*).

We break down patterns of AI usage in more detail in Figure 2. The median teacher uses the chatbot once per month on average and submits 3 queries per month; however, the top 10th percentile of teachers use it 5 days and submit 21 queries per month, as shown in Panel A.

Panel B shows a plot of retention, with the proportion of users submitting a query each month after being onboarded onto the system after a given number of months. 70% of onboarded users are active in the first month; this decays over time to roughly 25% active per month. Most usage occurs during school hours, but there is also substantial usage at night, suggesting it is used for planning, as shown in Panel C.

In an endline question, 40% of respondents reported sharing their phones with other teachers to allow them to use the chatbot, suggesting they found it valuable.

Heterogeneity

We assess heterogeneity in usage in Fig. 2 Panel D. Differences are not statistically significant, but we find suggestive patterns. The small number of teachers with a bachelor's degree use the assistant 130% more per month on average. This pattern of usage corroborates other work in low income countries that find AI assistance benefits higher performing entrepreneurs more (*11*), but contrasts with several studies in high income countries finding that AI assistance benefits lower skilled workers more (*13*, *14*). Classroom teachers use the chatbot slightly more than head teachers.[5] Women are slightly less likely to use the chatbot than men, echoing other studies (*15*). Teachers who used WhatsApp more frequently at baseline use the chatbot more. In contrast, it is the teachers who rarely use search at endline who use the chatbot more. This pattern would be consistent with the two technologies being substitutes, or with the chatbot working for teachers for whom search is less useful.

Threads

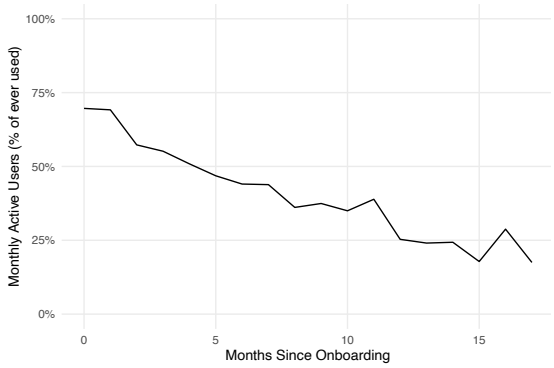
Some of the messages sent by teachers continue the existing conversation and some start new conversations. The WhatsApp interface does not have a button to start a new conversation, so we divide each user's interactions with the chatbot into separate threads, defined as conversations aimed at a common goal. We divide queries based on whether they referenced previous messages or the discussion, using a language model. We also count a message as starting a new conversation if it is separated by more than 48 hours. Based on this classification, 66% of queries start new conversations.

[5] We observe job title for the entire sample; for additional characteristics, we restrict to the survey sample.

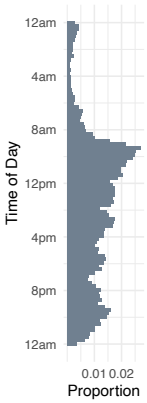
Panel A. Usage by teacher

	Mean	Q25	Q50	Q75	Q90
Active days per month	2.0	0.4	1.1	2.7	4.7
Queries per month	8.4	0.9	3.3	11.5	21.3
Queries	86.0	8.0	33.0	100.0	223.4
N	469				

Panel B. Retention



Panel C. Time of day



Panel D. Heterogeneity

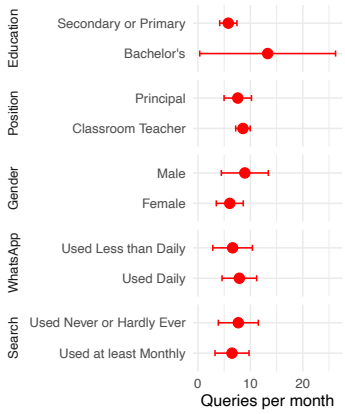


Fig. 2. Usage of AI chatbot. Panel (**A**) shows the distribution of usage by teacher, over the timespan from a teacher's first query to the end of the dataset, with one month computed as 30 days. Panel (**B**) shows that teachers continue to use the chatbot over time. Panel (**C**) shows usage by time of day. Panel (**D**) shows heterogeneity in usage. Position is measured for the entire sample; the remainder are measured in the surveyed sample (WhatsApp use at baseline; education and gender at midline; search at endline). Whiskers display 95% confidence intervals.

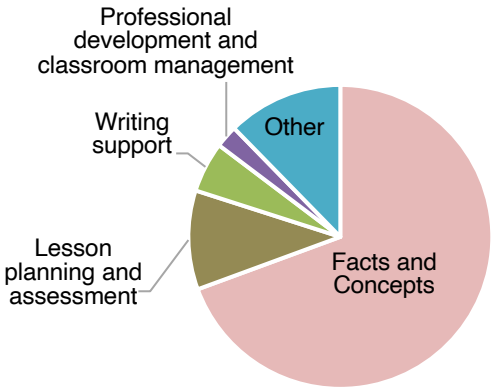
Tasks and Subjects

We next categorize the threads into tasks and subjects. We do this in a three step process. We first send the queries to a language model to suggest a list of categories. We manually edit this list to simplify and remove duplicates. We then send each thread to a language model and ask it to classify each one.

The largest task category is for facts and concept clarification (69%), followed by lesson planning and assessment (11%), writing support (5%), and professional development and classroom management (2%), as shown in Figure 3 Panel A. Thus, the largest category of queries represent retrieval requests that one might have expected search engines to be effective at answering.

We also apply a similar classification algorithm to the subjects of threads. These span the curriculum, as shown in Figure 3 Panel B.

Panel A. Tasks



Panel B. Subjects

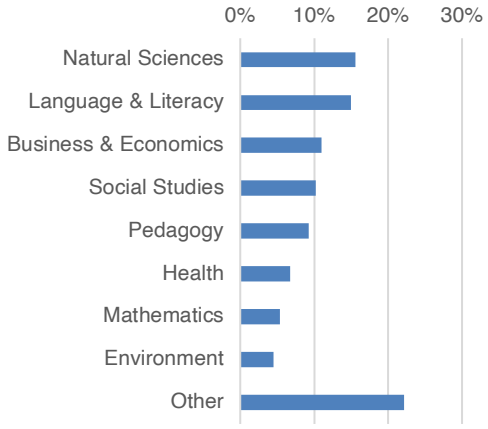


Fig. 3. Categories of chatbot queries. Classification of the threads submitted to the chatbot, by the type of task (Panel (**A**)) and subject (Panel (**B**)).

AI versus Search

We next consider how AI compares to search.

Self-assessment

The endline survey asked surveyed teachers how they determined when to use AI and when to use search; results are shown in Figure 4. When asked about times they use theTeacher.AI instead of search, all respondents reported at least one reason they did so; however, only 45% reported at least one reason when asked about times they used search instead of theTeacher.AI. The most common reasons to use AI were that chatbot gives a useful answer right away (85%), followed by that it is more concise (78%), more trustworthy (72%), creates content that cannot be found on the web (66%), or works better with the level of internet access (65%). Although 58% of teachers note AI's ability to ask follow ups, this is only occasionally used in our data. The most common reason that a respondent would use search instead of AI is that the web has content that the chatbot does not (18%). As a free response to this question, 5 users (3.4%) reported that since receiving access, they have stopped using web search and now instead use theTeacher.AI.

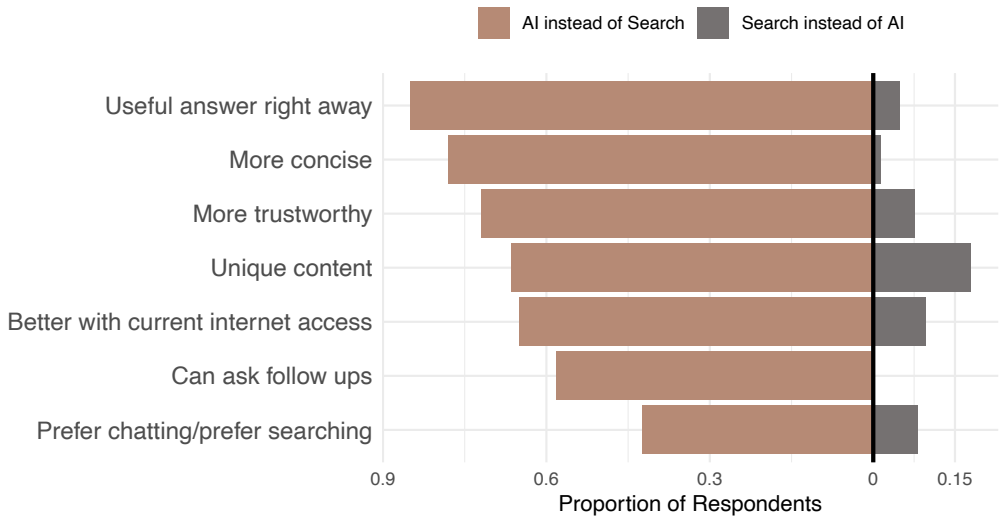


Fig. 4. Reported reasons for preferring AI versus search. Survey responses to the question, "When you use theTeacher.AI instead of internet web search for supporting your teaching activities, what are the reasons?" and the reverse question ("…internet web search instead of Teacher.AI…"). Multiple responses allowed. 'Other' response omitted, as some responded with general comments.

AI versus search results

We compare AI and search more systematically by analyzing the queries themselves. We take the queries submitted to the chatbot, and submit them to the local version of Google (google.com.sl), which is estimated to have 97% of search engine market share in the country (*16*). We use several settings to mimic the results that would be seen on a mobile phone physically located in Sierra Leone. We use a distributed virtual private network (VPN) service, ensure that associated IP addresses are geolocated in Sierra Leone, and mimic the browser characteristics of a lower-priced Android mobile device common in sub-Saharan Africa.

We consider only the first queries asked in each thread to give both technologies similar context, since otherwise AI would respond to natural language follow ups differently than search. We omit queries that are 'fillers', such as introductions or greetings. We compute the following statistics for a 12.5% random subsample of queries.

Very few of the search results are local content. Only 2.2% of returned web sites are at a Sierra Leonean domain (.sl), and few are from similar countries at all: 86% are US domains (.com, .org, .edu, .net, .gov) and 3% are UK domains, as shown in Figure 5. A breakdown by domain is given in Table S1.

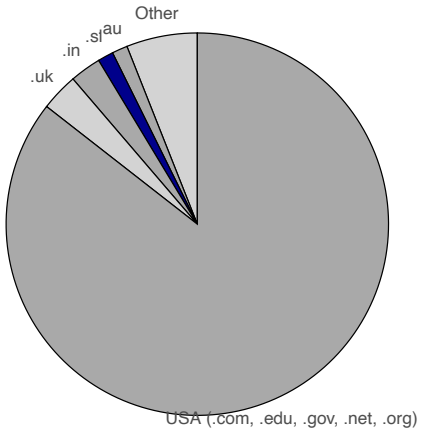


Fig. 5. Origin domain of search results for Sierra Leonean teacher queries. The top-level domain (or country code, for non-U.S. domains) of the top 5 search results from Google.com.sl for queries submitted to the chatbot. (Sierra Leone domains emphasized in dark blue.)

We then compare the first 5 search results to the chatbot response based on format and content. We obtain organic results, omitting special result sections, such as maps, images, or videos. We then download each search result and extract its main text. Standard web pages include text that is not part of the main content (such as navigation links), so for these we extract only the main article using an automated method to remove extraneous content. For PDFs, we convert the entire file to text. We also compute the bandwidth needed to load each result by loading a random subset in a virtual browser. We omit search results that have main content that is video rather than text (the 1.4% of results pointing to YouTube), as well as from a set of domains that commonly block content behind a paywall or login form (4.4% of results).[6] We also omit 17.3% of search results for which we encountered scraping difficulties. Three example queries and responses are given in Tables S2-S4.

Format

AI responses are shorter by an order of magnitude (mean word count of 156 versus 3,311).[7] Both because AI responses are shorter, and because they are text (without syntax, styles, images, ads, and code), they require transmitting much less data, as shown in Figure 6. The average data needed to transfer an AI response is 0.80 kilobytes (KB), relative to 2,499 KB for a search result. Thus, a search result requires 3,107 times more data to load than an AI response on average. The data used by an individual search result is in addition to the data needed to transfer for the list of search results in the first place. Larger file sizes not only incur larger transfer costs, they are also challenging to download on slow and intermittent networks.

[6] We classify results as behind a login form for jstor.org, study.com, brainly.com, tiger-algebra.com, and reddit.com. Content at these sites could not be accessed by our scraper, and may only be accessible to teachers who register.
[7] Note that some AI responses are broken into chunks, so there is a natural limit to how long they can get; a user may reply "Continue" to see the following chunk.

Overall, these suggest that relative to chatbot responses, web content is poorly formatted to be viewed on the small devices and low bandwidth connections used by our sample of teachers. The web does possess high quality content; for example, 1.6% of search results are from the Sierra Leonean education ministry (representing 74% of all results from Sierra Leone), including localized lesson plans and syllabi. But among those, 92% are PDF files, some over a hundred pages long and difficult to view on a small screen.

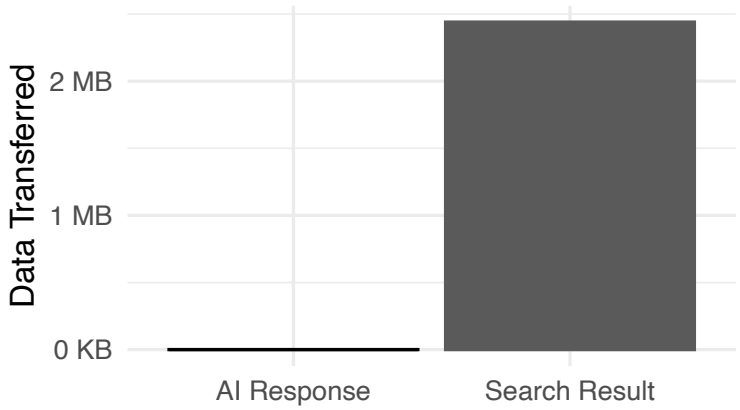


Fig. 6. Data usage of average AI response and web search result. The average search result consumes 3,107 times more data to load.

Cost

The number one reported limiter that prevents individuals in sub-Saharan African countries from using the internet more is cost (*5*), as shown in Figure S3. Web pages use much more mobile data to download, but AI responses cost more to compute. For a given query q, we compute the cost of loading a single search result as

$$c_{search}(q) = p_{transfer} GB(x_{search}(q))$$

and the cost of an AI response as

$$c_{AI}(q) = p_{transfer} GB\big(x_{AI}(q)\big) + p_{compute,in} Tokens(S + q) + p_{compute,out} Tokens\big(x_{AI}(q)\big),$$

where $x_j(\cdot)$ represents the content retrieved by method j in response to a query, $GB(\cdot)$ counts the bandwidth required to transmit that content, $Tokens(\cdot)$ computes the amount of AI compute required for the query and system message, and S represents the total tokens used for the system message.[8] Prices include the cost per gigabyte of data transfer ($p_{transfer}$), per AI input token ($p_{compute,in}$), and per AI output token ($p_{compute,out}$).

[8] A system message is included both at the start and end of the query, which can make the system more secure, as described in Appendix B1. Our main cost computations include $S = 2s$, where include s is the token count of a single system message.

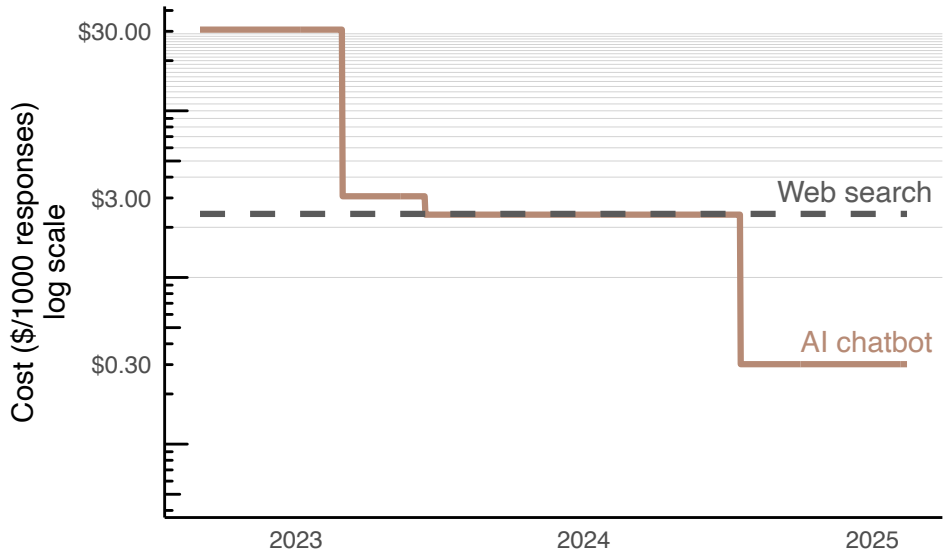


Fig. 7. Cost comparison over time. The cost per query is plotted over time, for the average AI chatbot response and search result, in log scale. Costs include mobile data transfer in Sierra Leone of $p_{transfer} = \$1.01$ per gigabyte (*17*), and, for AI, the compute cost for OpenAI GPT models (GPT3 Davinci until Nov 2022, GPT3.5 Turbo until July 2024, then GPT4o mini) applied to the number of tokens used for the query and model output.

We plot a cost comparison in Figure 7 for the average search and AI response. Given a mobile network cost of $1.01 per gigabyte (17), search on the Sierra Leonean mobile network cost $2.41 per thousand results during this period (dashed gray line). AI responses (salmon colored line) cost $30.70 per thousand results in early 2022. However, the cost of AI has declined precipitously, to $2.38 in 2023, and to $0.30 per thousand results in 2024. These cost comparisons suggest it is already 8 times more cost effective to use AI than web search. If the steep downward trends in AI computation costs underlying Fig. 7 continue, AI will become even more cost effective. This calculation omits the bandwidth required to transmit the search result page itself, and only considers the bandwidth to load a single search result: a search interaction that compares multiple results would make search cost even more. We do not factor in follow up queries for either technology. We present costs under alternate assumptions in Table S8. Search retrieves content from the web and delivers it verbatim, this comparison suggests that in this context it is far more efficient to tailor knowledge on the cloud and deliver only what the user needs.

Content

Search results contain similar information as AI responses. We assess this by sending each query, AI response, and search result to a large language model. We prompt the model to assess the degree to which the text of the search result contains the main ideas in the AI response, on a scale from 1-10, where 1 indicates that the search result has no core ideas of the AI response, and 10 indicates that it captures all core ideas. We find that 46% of top 5 search results contain most of the core ideas of the AI response (score 6 or above), and for 74% of queries there is a top 5 search result that contains most of the core ideas of the AI response. We additionally assess

whether the query itself is conducive to a single correct answer (like a factual question) or multiple possible correct answers (like asking to generate a story) by asking a language model to classify it. By that classification, 48% of queries are conducive to a single correct answer. We plot the distribution of these two variables together in Figure S5 and S6, showing that contains scores are higher for queries rated as having a single correct answer. Overall, these results suggest users are mostly asking the AI system to retrieve information, not generate new content.

Quality

We next compare the quality of the responses. We recruit 25 teachers from around the world to evaluate them. We draw a random sample of 100 queries, restricting to those for which search results were successfully scraped. For each, we assembled a dataset with the query, the AI response, and extracted text from one of the top 5 search results (selected at random).[9]

Respondents were recruited on the Prolific survey platform from a pool of teachers. Each was shown a sequence of 10 queries randomly selected from the 100. Alongside each query, they were shown a single response, and asked to rate it along various dimensions. The response was randomly selected to be either the AI response (2/3 probability) or the extracted search result text (1/3 probability). We oversampled AI responses because they were shorter and thus faster to evaluate, and because we wanted to obtain more precise measures of their factuality. To avoid bias, respondents were blinded as to the origin of the responses, both in general, and for each query. They were informed that in rating the response they could use their own knowledge or consult outside resources, such as books or the web, but were asked not to consult an AI chatbot.

We present results in Figure 8. On a five-point scale, respondents rate the AI response as more relevant (+1.23, p-value < 0.001) and helpful (+1.12, p-value < 0.001).

[9] We omitted some queries for confidentiality and due to data quality. We omitted 3 queries that named people, places, or organizations in Sierra Leone, for confidentiality reasons. We omitted 2 queries that were improperly threaded (they continued a previous conversation). We omitted 3 queries for which the randomly selected search result did not have content (e.g., it was a paywall or otherwise). We also omitted 3 queries that the chatbot classified as not related to education, and thus did not provide a response. We replaced each omitted query with a randomly sampled query, to reach a final sample of 100.

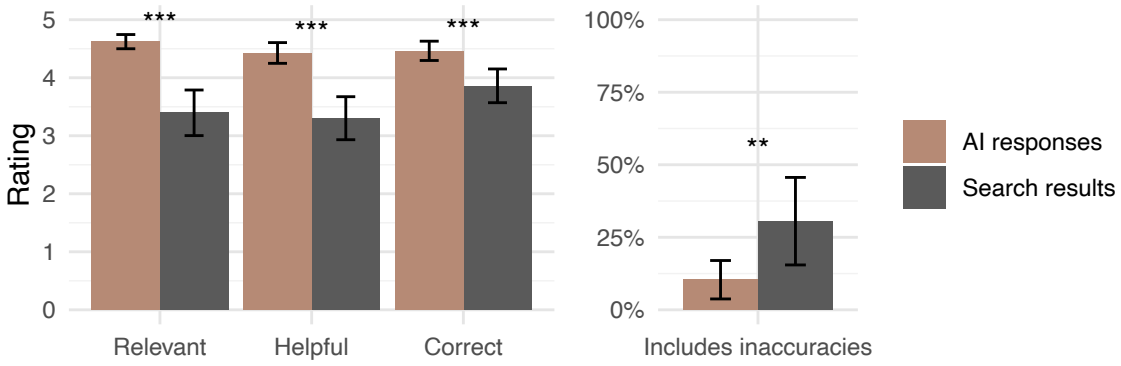


Fig. 8. Independent teacher ratings of AI responses and web search results. For a sample of 100 randomly selected queries, with associated AI response and search result randomly drawn from top 5 results. 95% confidence intervals shown with whiskers; standard errors clustered by query and respondent. Stars represent significance level of a two-sided F-test of equality; ***: p<0.01, **: p<0.05.

There is a general concern that AI systems may generate responses that are not correct ('hallucinate'), either not internally consistent with the instructions given, or not factually consistent with the world (*18*, *19*). The random sample of 100 AI responses includes at least three hallucinations that we verify independently.[10] However, respondents rate fewer AI responses as containing inaccurate information (10% vs 31%, p-value of difference=0.023), and on average containing more knowledgeable, correct information (difference +0.6, p-value < 0.001, on a 5-point scale).

Why do respondents rate the search results as having more inaccuracies? Respondents were asked to describe their reasoning in a free response question. Some explain that they considered a search result inaccurate because it made unfounded claims. Other results did not directly respond to the query, included irrelevant information, or did not completely answer the query. For example, a query about 'effect' returned a result about 'affect.' In analyzing these results, it became clear that typically when using search, one does not expect individual results to necessarily be accurate, or to answer the question, but rather requires the user to compare the veracity and relevance of multiple sources. But that comparison is challenging on small screens and requires even more data than loading a single response. Our experimental structure that blinds the respondents to the source of the response makes clear that individual search results often fall short when viewed as answers to questions.

We interpret the positive ratings of correctness to suggest that the information teachers are asking for is generally well known by the system, for which hallucinations are less common. Corroborating this, in the endline survey of teachers using the system in Sierra Leone, 72% of teachers reported that when they use AI rather than search a reason is they found AI more trustworthy. Yet it is well known that AI can make mistakes, so an interface that conveys a

[10] One claims that there are no 2 letter words in English before listing some in the next sentence; one claims that the Dead Sea is the saltiest body of water on Earth (Wikipedia suggests otherwise); one claims that a simple Excel formula is incorrect. The newer model GPT 4o mini answers these queries correctly.

single definitive answer can be misleading. This is particularly important for users who do not have external resources with which to validate responses.

Discussion

Our results show that the web has profound gaps in low-resource settings. With web search, results are limited to existing content in its existing format. There is little such content tailored for low-resource communities. In contrast, AI can tailor a response, and 'interpolate' between preexisting pieces of content. This property makes it possible to provide existing knowledge in more appropriate formats, which can be several orders of magnitude more data-efficient. But it may also generate new content. In our data, we see teachers ask the chatbot to generate stories that have specific characters and morals and are localized to Sierra Leone, content that does not exist on the web. This property of AI can be more useful when content is scarce, such as is the case in much of sub-Saharan Africa. Thus, even if there exists less digital data on the lives and needs of low income communities on which to train AI models (*20*, *21*), AI may be able to extract more value out of what data there exists than previous retrieval methods.

The tested chatbot is simple: at this time, it was using an earlier model (GPT 3.5 Turbo) and a prompt that references that teachers have low resources but does not reference Sierra Leone. More advanced models are known to reduce hallucinations, and more tailored prompts and fine-tuning with local data are likely to improve usefulness. Thus, this study should be thought of as illustrating a lower bound for the potential of these models.

When considering broader implications, three caveats should be made. First, this is a population that speaks English, which is the language in which current language models perform best. AI responses are not yet as useful in lower resource languages. Second, teachers use the system as one of many inputs into classroom decisions, and are skilled at considering the veracity of responses. Caution is warranted when allowing AI to directly make high stakes decisions without such review. Third, education may be a complement to using these systems. We found suggestive evidence that teachers with higher education used the system more frequently. During pilots of the system, some teachers struggled to craft queries that the chatbot could usefully answer. Training was helpful. Thus, current AI systems may complement some existing human capital; more refined designs may be necessary for users with less education.

Despite these caveats, relatively high rates of chatbot usage in low-income countries is not unique to our study. A study of web traffic found that lower middle and lower income countries account for only approximately 15% of search traffic, but approximately 25% of AI chatbot web traffic (*22*). Our results suggest that AI can already deliver information at one eighth of the cost of the web, which has potentially profound implications for access to information. While much discussion worldwide is about whether and how AI will trigger a revolution of intelligence, this study suggests that one impact of the technology could be to partly catch up low-income countries to the revolution of information.

Methods

Recruitment

Fab Inc recruited teachers through public schools and through a private school network (EducAid Sierra Leone). The chatbot works only for users explicitly given access. During onboarding, teachers were given a training brochure. The brochure listed possible uses of the system, including for factual questions, generating content, and getting advice (Figure S8a-c). Teachers were informed that conversations with the chatbot would be analyzed, and were asked to not identify themselves or provide confidential information. One cohort, which we call the survey sample, was surveyed three times by EducAid. They were surveyed during onboarding ('baseline', in winter 2024), 2-3 months after onboarding ('midline'), and 3-4 months after onboarding ('endline').

Data

The chatbot backend records each query submitted by a teacher, a pseudonymous identifier for each teacher, a timestamp, and the corresponding response provided by the AI system. Fab worked with EducAid to develop survey questions covering AI awareness, teaching practices, job satisfaction, and demographics. To this, the Columbia University researchers added questions about web search usage. We restrict the sample to teachers and principals in the public schools or EducAid, with Sierra Leonean phone numbers, whose first query was on a known day of teacher training. This omits some users who were given access for testing purposes or other channels. We omit queries from each user's first day, to omit any practice questions submitted during the training session, and include only users who successfully submitted at least one query after the training day.[11] The survey sample includes N=166 teachers surveyed at baseline, of which N=122 could be linked to the query data. That declines in the midline (N=152 surveyed, N=113 linked) and endline (N=146 surveyed, N=109 linked). Fab Inc. provided the Columbia research team access to usage and survey data under an agreement that gives the academic research team freedom over publication of results, so long as it does not violate confidentiality or privacy rules.

Analysis

We assembled the main dataset using a multistep process. Figure S7 shows a diagram of major steps. We report validation exercises for each step in Appendix A. For steps that use a large language model (LLM), we submitted these to OpenAI GPT-4o-mini (`gpt-4o-mini-2024-07-18`), using the prompts described in Appendix B. Steps are as follows:

a. Categorizing queries into threads. TheTeacher.AI does not have a "new conversation" button, so new conversations are not automatically grouped into separate threads. We used an LLM to categorize queries and AI responses into threads (a conversation aimed at a common goal). For each user, we processed the queries and the chatbot's responses sequentially. Given a current proposed thread, we assessed whether a subsequent query-response pair should be included in the ongoing thread based on whether the new message continued the conversation, referenced previous messages, requested further clarification, or introduced new information relevant to the ongoing discussion. Queries separated by more than 48 hours were considered to

[11] The full dataset includes N=1,038 unique users. The sample falls when adding restrictions: teachers/principals (N=834); in Sierra Leone (N=765); in public schools or EducAid (N=762); starting on known training dates (N=529); and successfully submitting at least one query after the training day (N=469).

belong to different threads, regardless of the other criteria. If the new message was determined to belong to the currently-examined thread, it was appended to the current thread. Otherwise, the said current thread was closed, and the new message was treated as the start of a new thread. This process was repeated for the entire sequence of messages to classify them into distinct threads. The LLM prompt is given in Appendix B2.

b. Categorizing threads into tasks and subjects. We classify unique threads by task and subject. We initially asked an LLM to summarize each query, and then generate a list of tasks that classify the intent of the teacher based on the queries. We manually edited this list of task categories, and then ask an LLM to classify each thread, from among the following list:

- Concept Clarification and Factual Information
- Lesson Planning
- Writing Support
- Teacher Professional Development
- Classroom Communication
- Student Behavior Management
- Exam and Assessment
- Parent and Community Engagement
- Greetings, Gratitude, and Miscellaneous Queries
- Supervision and Collaboration with Other Teachers

We repeat this exercise for a subject label, aiming to classify the discipline or field the teacher is querying about. We ask the LLM to classify threads among the following categories:

- Business & Economics
- Language & Literacy Development
- Social Sciences & Social Equity
- Natural Sciences
- Mathematics
- Pedagogy & Instruction
- Health, Wellness, & Physical Education
- Environmental Studies & Sustainability
- Technology & Engineering
- History & Cultural Studies
- Teacher's Own Personal Development & Life Skills
- Arts & Humanities
- Community Involvement & Educational Partnerships
- Behavioral Management & Classroom Strategies
- Ethics & Values in Education

Classification uses the prompt in Appendix B3.

c. Determining queries' status as being 'filler'. We classified a query as filler if it was merely an introduction, greeting, thank-you message to the AI assistant, a wrap-up of the conversation, or if it consisted only of numbers (e.g., "2" or "44") without further explanation, or short responses like "yes" or "no." However, if a query involved asking for further information or

making a non-trivial statement, it was not classified as filler. To assess this, we passed batches of queries to an LLM with the filler prompt given in Appendix B4.

d. Gathering search engine results. We next submitted queries to Google search to gather the web search results that it would have led to. We consider only queries that are the first in a thread, as follow up questions could have resulted in chatbot responses being more tailored than an equivalent search. We also omit filler queries. To get search results as close as possible to what a teacher in Sierra Leone might see, we used a proxy service, Oxylabs, that allows one to rent the use of internet protocol (IP) addresses in a given location. We request IP addresses located in Freetown, Sierra Leone. We initialize a web browser with settings that mimic a mobile phone (using Chrome Webdriver and Selenium). We emulate a common brand of phone in sub-Saharan Africa, by using the Tecno Spark 8C phone's screen size dimensions and resolution. With our web browser we first verify that the IP address is located in Sierra Leone before proceeding, using the service IPinfo. We then submit the query to google.com.sl. Once we submit the query, we retrieve the individual search result URLs that show up on the webpage. We omit any special sections of the search results page, such as videos, images, and shopping.

e. Scraping content of web pages. We scraped the top 5 search results for each query to capture data usage and content. We use Selenium to request the content of each URL, and we download the content. We use several settings to encourage web pages to localize their responses to a typical mobile phone in Sierra Leone. We set the user agent to simulate requests from a Tecno Spark 8C phone, prioritize Sierra Leonean English (with the header field accept-language: en-SL, en), and set the save-data field to on to encourage data efficient responses. For PDF files, we convert the entire document to text. For standard web pages (in Hypertext Markup, or HTML), we extract main text using the library Trafilatura, which reports good performance evaluations (https://trafilatura.readthedocs.io). For some pages that were returned, the resulting document did not contain content (it may have been blocked by a captcha or paywall). We thus classify whether a resulting document represents a scraping error, using an LLM with the prompt given in Appendix B5.

f. Data usage of web pages. For a random subset of 1,000 queries, we carefully measured the data consumption of each of the top 5 search results. We use Selenium with the above settings, plus additional settings allowing the server to compress transmissions using standard protocols (gzip, deflate, or brotli), and disabling the cache. To assess the data transfer required for the page along with its embedded resources (such as images, fonts, styles, and code), we monitor logs of network transmissions as the page loads. To ensure the page is fully loaded, after the body is loaded, we wait 15 seconds and then add additional waits of 2 seconds up to 5 times if the page continues interacting with the network. Validations suggest our data use measure for search results is conservative. For PDF files we monitor the size of the download.

g. Costs. We compute the number of tokens in text using the OpenAI tiktoken library, using the tokenizer associated with gpt-3.5-turbo, and compute costs using OpenAI's posted prices, as

displayed in Table S7. We compute the bytes necessary to transfer a string of text using the UTF-8 encoding that is most used on the web.

h. Automated content comparison. For each query, we then compared the content of the AI response to the article content of the top 5 search results. We computed several metrics:

- *Contains*: Whether the ideas of the AI response are contained within the article was assessed by prompting an LLM to capture the core ideas in the AI response and detect how many of those core ideas are present in the article. The LLM classifies the article with a score between 1-10 describing the how many core concepts it contains from the AI response. If the article is larger than 2000 characters, we first summarize it with an LLM. Prompts are given in Appendix B6.

- *Answer Uniqueness*: We classified queries based on how many potential answers they were likely to have (1 answer, 2-10 answers, 11-500 answers, 501-Infinite answers, and Uncertain). Prompt given in Appendix B7. In some analyses we simplify by collapsing categories to one or many answers.

i. Human content comparison. We recruited an independent sample of 25 participants from Prolific.com between January 5-7, 2025 to rate responses. To be eligible, participants had to identify as being a teacher. The study was given a generic title ('Information for teachers') and description ('In this study you will rate different responses to a query') to avoid priming participants. Participants were recruited from the global sample around the world and were unlikely to be users of the chatbot system.

Practice. The first section of the survey presented two example queries and responses, and asked for practice ratings. Participants were first presented with the query, "How can I teach fractions to 6th grade students?", and asked to rate the relevance of a(n irrelevant) response. Then teachers were presented with the query, "What is important for students to know about Napoleon," and asked to rate the correctness and factuality of a (hallucinated) response to the question, which sounded plausible but included many falsehoods. Participants were allowed to continue when they rated the first response as lacking relevance and the second as having incorrect information and made-up facts.

Main survey questions. The main section of the survey asked participants to rate responses to 10 different queries. For each pair of a query and response, we asked the questions,

1. Relevance: Does the response make sense in the context of the question? (1-5 rating)
2. Helpful: How helpful is the response to the question? (1-5 rating)
3. Correctness: Does the response contain some knowledgeable, correct information? (1-5 rating)
4. Inaccuracies: Did the response contain any inaccurate or fabricated information? (Yes/No/Uncertain/Not applicable)
5. Please explain why you rated the correctness and factuality of the response as you did. What evidence or resources did you rely on to reach your conclusion? (Open response)

At no point did we mention the origin of the responses. The final section of the survey asked for demographics.

Data and materials availability: The core chatbot code is available in an open source repository at https://github.com/AI-for-Education/fabdata-llm, and all relevant language model prompts are included as appendices. Analysis code and survey data will be made available. Because the queries submitted by individual users are confidential, they can only be shared with a data sharing agreement, which are available upon request.

References and Notes

1. GSMA, The State of Mobile Internet Connectivity: Sub-Saharan Africa key trends (2023). https://www.gsma.com/r/wp-content/uploads/2023/10/State-of-Mobile-Internet-Connectivity-2023-Sub-Saharan-Africa.pdf.
2. International Telecommunication Union, ITU DataHub (2024). https://datahub.itu.int/.
3. R. Jensen, The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector. *Quarterly Journal of Economics* **122**, 879–924 (2007).
4. J. Hjort, J. Poulsen, The Arrival of Fast Internet and Employment in Africa. *American Economic Review* **109**, 1032–1079 (2019).
5. ResearchICTAfrica, "After Access Survey" (2022).
6. A. Ballatore, M. Graham, S. Sen, Digital Hegemonies: The Localness of Search Engine Results. *Annals of the American Association of Geographers* **107**, 1194–1215 (2017).
7. Afrobarometer, version Round 10 (2024); http://www.afrobarometer.org.
8. J. H. Choi, O. Garrod, P. Atherton, A. Joyce-Gibbons, M. Mason-Sesay, D. Björkegren, "Are LLMs Useful in the Poorest Schools? TheTeacher.AI in Sierra Leone" in *NeurIPS Workshop on Generative AI for Education* (2023; http://arxiv.org/abs/2310.02982).
9. R. Abebe, S. Hill, J. W. Vaughan, P. M. Small, H. A. Schwartz, Using Search Queries to Understand Health Information Needs in Africa. *Proceedings of the International AAAI Conference on Web and Social Media* **13**, 3–14 (2019).
10. J. Mullan, A. Taddese, "EdTech in Sierra Leone: A Rapid Scan" (3, EdTech Hub, 2020); https://doi.org/10.53832/edtechhub.0038.
11. N. G. Otis, R. Clarke, S. Delecourt, D. Holtz, R. Koning, The Uneven Impact of Generative AI on Entrepreneurial Performance. doi: 10.31219/osf.io/hdjpk (2024).
12. Education Endowment Foundation, Evidence and resources, *EEF* (2024). https://educationendowmentfoundation.org.uk/education-evidence.
13. E. Brynjolfsson, D. Li, L. R. Raymond, Generative AI at Work. National Bureau of Economic Research 31161 [Preprint] (2023). https://doi.org/10.3386/w31161.
14. S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer, The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv arXiv:2302.06590 [Preprint] (2023). https://doi.org/10.48550/arXiv.2302.06590.
15. A. Humlum, E. Vestergaard, The Adoption of ChatGPT. 4827166 [Preprint] (2024). https://doi.org/10.2139/ssrn.4827166.
16. StatCounter, Search Engine Market Share Sierra Leone, *StatCounter Global Stats*. https://gs.statcounter.com/search-engine-market-share/all/sierra-leone/2022.
17. ResearchICTAfrica, RIA African Mobile Pricing Index, *Research ICT Africa* (2024). https://researchictafrica.net/research/ramp-index/.
18. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 248:1-248:38 (2023).

19. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, doi: 10.1145/3703155 (2024).
20. D. Björkegren, Artificial Intelligence for the Poor, *Foreign Affairs* (2023). https://www.foreignaffairs.com/world/artificial-intelligence-poor.
21. S. Mirza, B. Coelho, Y. Cui, C. Pöpper, D. McCoy, Global-Liar: Factuality of LLMs over Time and Geographic Regions. arXiv arXiv:2401.17839 [Preprint] (2024). https://doi.org/10.48550/arXiv.2401.17839.
22. Y. Liu, H. Wang, Who on Earth Is Using Generative AI ? doi: 10.1596/1813-9450-10870 (2024).

Acknowledgments: Thank you to Gavin Dhanda, Corey Gelb-Bicknell, and Joachim Rillo for excellent research assistance. This project benefited greatly from conversations with María José Ogando Portela at Fab Inc.; Andrew Joyce-Gibbons at the Curriculum Foundation; Miriam Mason-Sesay at EducAid Sierra Leone; as well as with Jeff Berens and Siddharth Suri. Fab Inc. provided access to data on queries and surveys of Sierra Leonean teachers. Columbia University IRB provided human subjects oversight of the research. The Columbia University team retained full intellectual freedom in conducting this research, and as such all opinions and errors are ours.

Inclusion & Ethics: Query data gathered in Sierra Leone was collected by the organization Fab Inc, which works closely with local stakeholders. Sierra Leonean survey data was collected jointly by Fab Inc and EducAid Sierra Leone.

Funding: Columbia University School of International and Public Affairs Faculty Grant funded the research project.

Author contributions: D.B. conceived and supervised the research project, and wrote the paper. D.B., J.C., D.P.B., and D.S. analyzed the data, prepared figures, and edited the paper. P.A. and O.G implemented the chatbot in Sierra Leone, coordinated activities with partner EducAid and the Ministry of Basic and Senior Secondary Education (MBSSE) in Sierra Leone, and provided data to the Columbia University team.

Competing interests: D.B., J.C., D.P.B., and D.S. declare that they have no competing interests. P.A. and O.G. originally developed the chatbot at Fab Inc. as part of their broader consultancy support work to MBSSE and are shareholders in Fab Inc. They have since spin-off the chatbot implementation into a not-for-profit organization, AI-for-Education.org, which coordinates this and other AI work in low- and middle-income countries through grant funding.

Supplementary Information

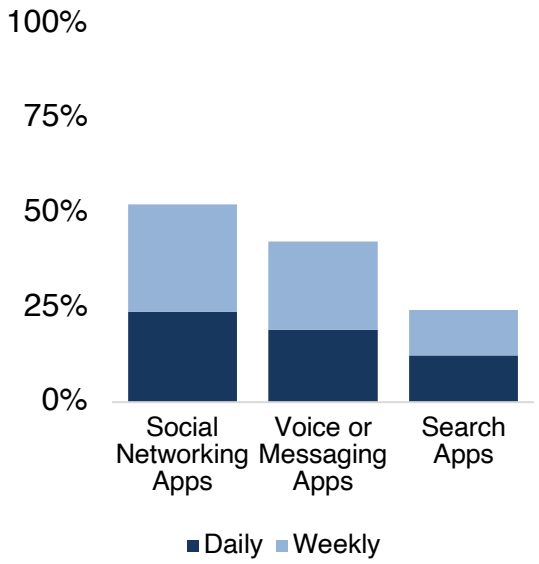
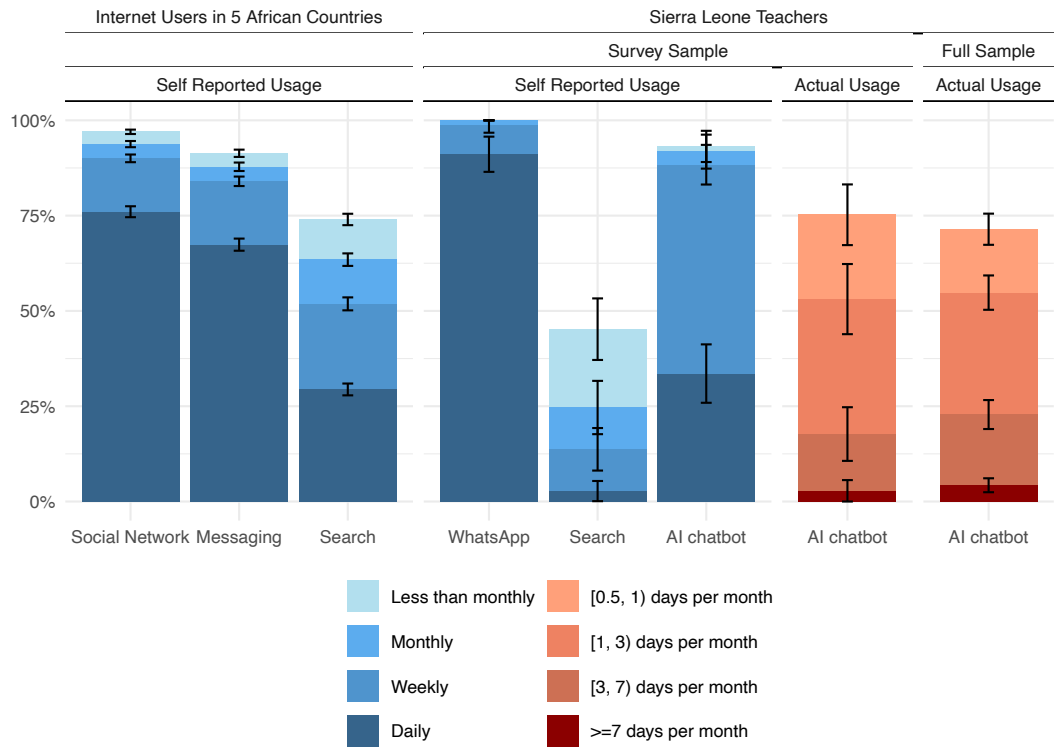


Fig. S1. App usage among Feature/Smartphone Owners in 10 African Countries (2017). Survey covering Ghana, Kenya, Lesotho, Nigeria, Mozambique, Rwanda, Senegal, South Africa, Tanzania and Uganda using national representative survey weights

Panel A. All Usage Measures



Panel B. Where do you find the information to learn more about teaching? (endline)

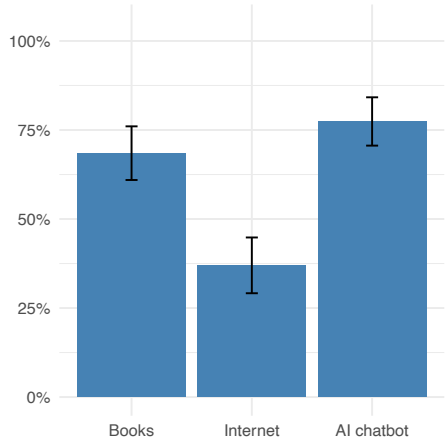


Fig. S2. Alternate measures of usage. Panel (**A**) shows self-reported usage of various services. We also show the actual usage from queries based on average days active per 30 days, for both the surveyed subsample and the full sample. In Sierra Leone teacher sample, self reported categories are 'Every day', 'Several times a week', 'A few times a month', 'Hardly ever', and 'Never' (omitted). Panel (**B**) shows resources that teachers report consulting in endline survey. Error bars indicate the 95% confidence intervals for that frequency or higher.

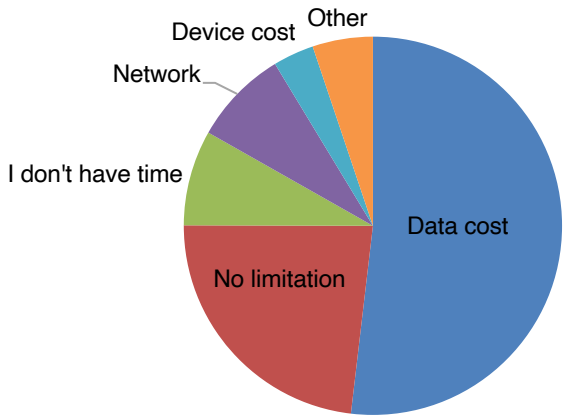


Fig. S3. Limiters from using the internet in 5 African countries. Response to 'What is the main issue limiting you from using the internet more often?' among internet users in Ethiopia, Kenya, Nigeria, South Africa, and Uganda in 2022. (*5*)

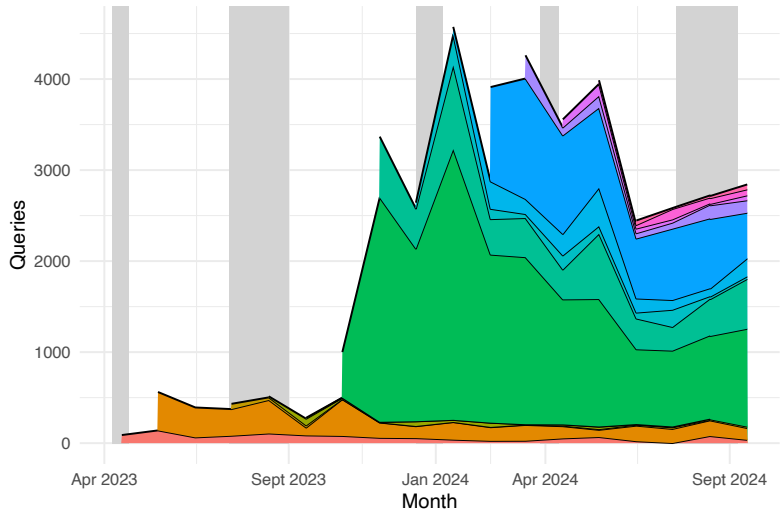


Fig. S4. Usage over time. The number of queries submitted by enrolled teachers per month. The plot shows the onboarding of multiple cohorts, with each cohort shown with a horizontal ribbon of a different color. School breaks are shaded in vertical gray bands. Usage remains sustained in the months that follow, though within each cohort there is a slight decline over time. Usage tends to drop towards the end of the school year in June, and tend to rise prior to the start of school in September when teachers are planning for the new year.

Panel A. Distribution of Contains Scores: Does Search Result Contain Ideas of AI Response?



Panel B. Distribution of Contains Scores: Does Search Result Contain Ideas of AI Response?

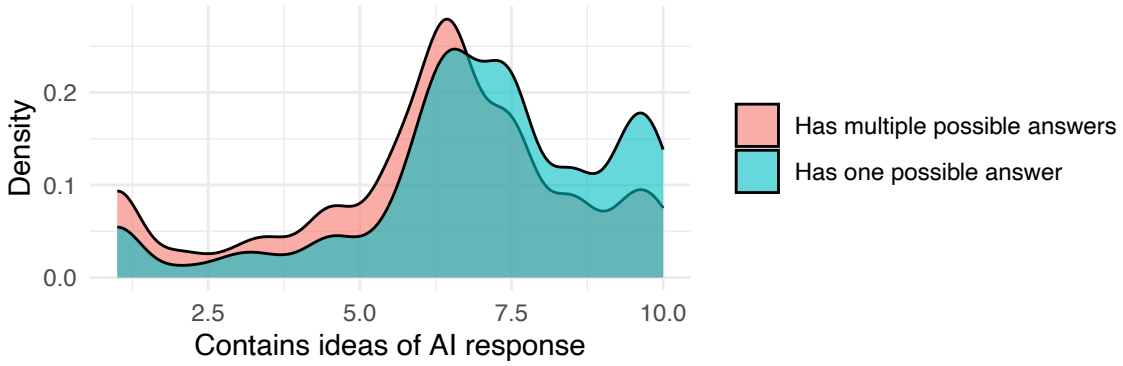


Fig. S5. Do Search Results and AI Responses Contain the Same Ideas? Panel (**A**) shows the distribution of contains scores: a rating of the degree to which the search result contains the main ideas of the AI response for the same query, as assessed by a language model with prompt given in Appendix B6. The search result with the highest contains score for each query is plotted. Panel (**B**) splits this distribution by whether the query itself is rated to have one possible answer (such as a factual query) or multiple possible answers (such as a creative query), as assessed by a language model with prompt given in Appendix B7. Contains scores are higher for queries with one possible answer; for queries with multiple possible answers there is a bump at a score of 1: search results that contain none of the ideas of the AI response.

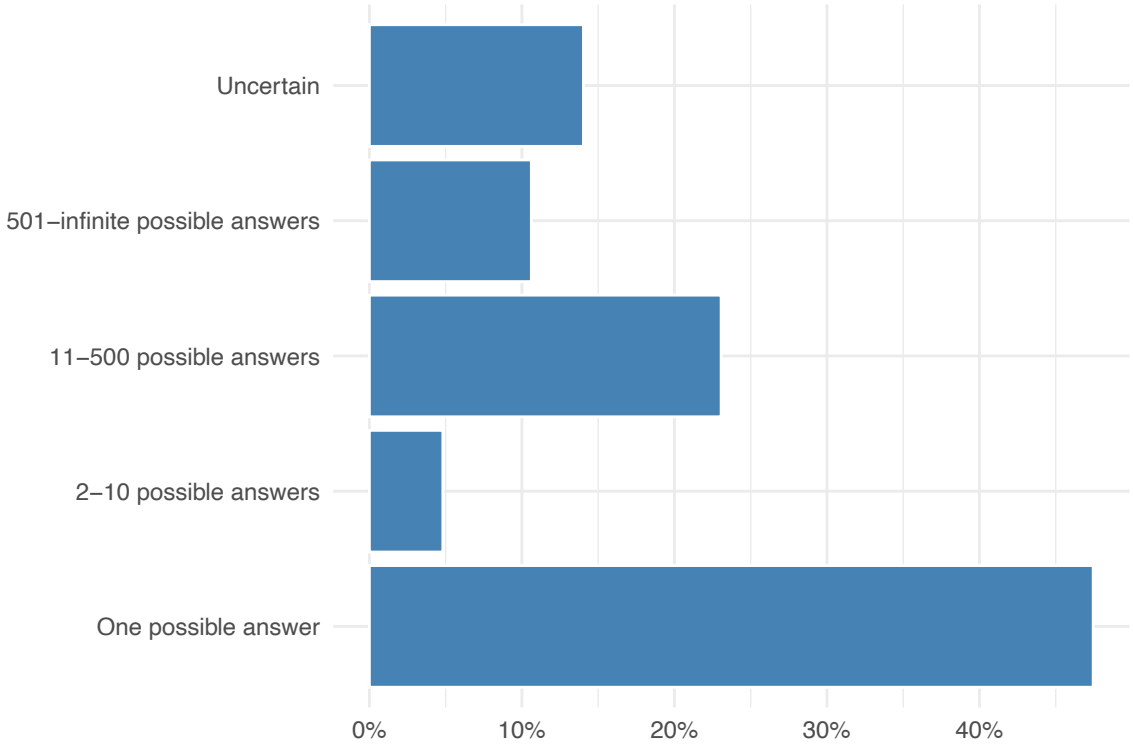


Fig. S6. How Many Possible Answers Do Queries Admit? Queries were rated based on approximately how many possible answers they admit (for example, a factual query may have one possible correct answer; a query to create a new story may have infinite possible answers). Queries were rated by a language model with prompt given in Appendix B7. 47% of queries are rated as having one possible answer.

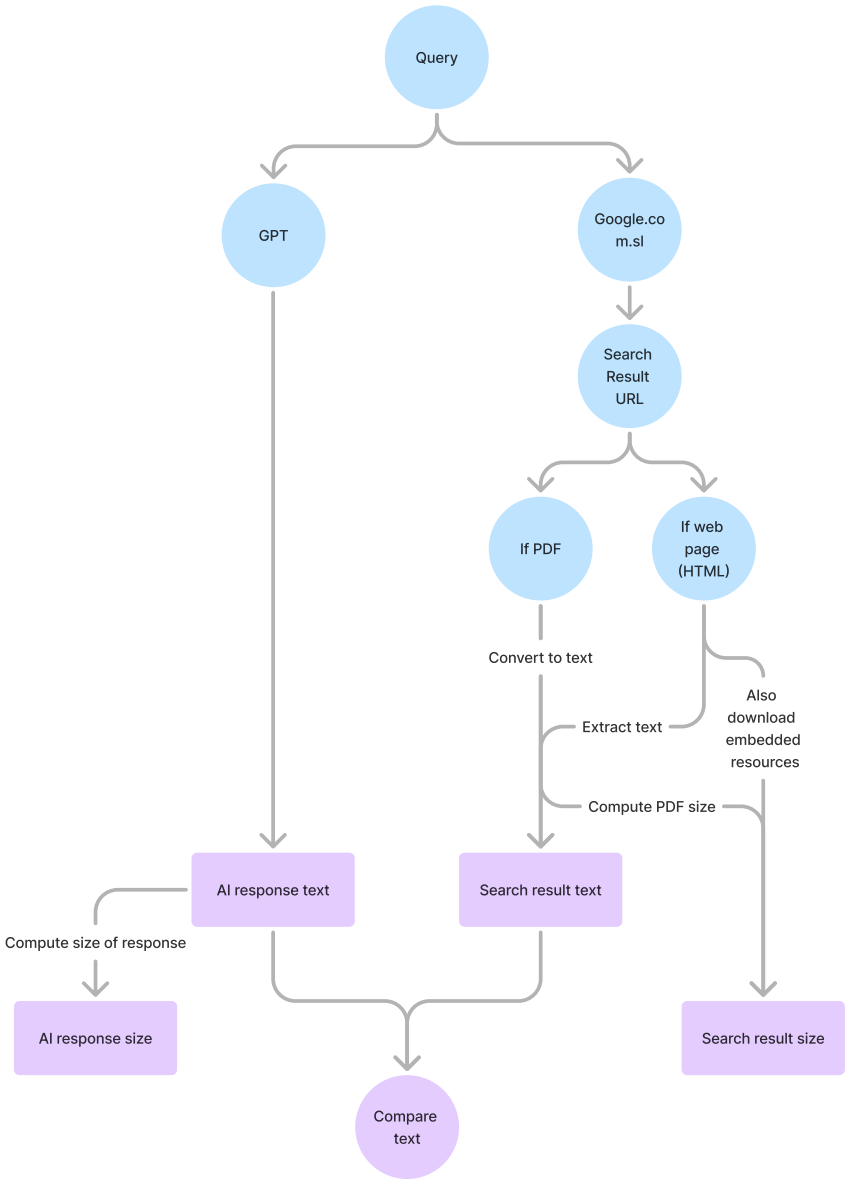


Fig. S7. Query Process Diagram. Overview of process for analyzing queries and responses. In the left branch, the AI response was generated by GPT for theTeacher.AI at the time the query was submitted. In the right branch, results were sent to Google.com.sl.

TheTeacher.ai is a chatbot on WhatsApp that can answer all sorts of questions for you to help with your teaching. The answer is not coming from a person. It is free for you. We hope you enjoy using it.

1. Take it in turns to read through the information below:



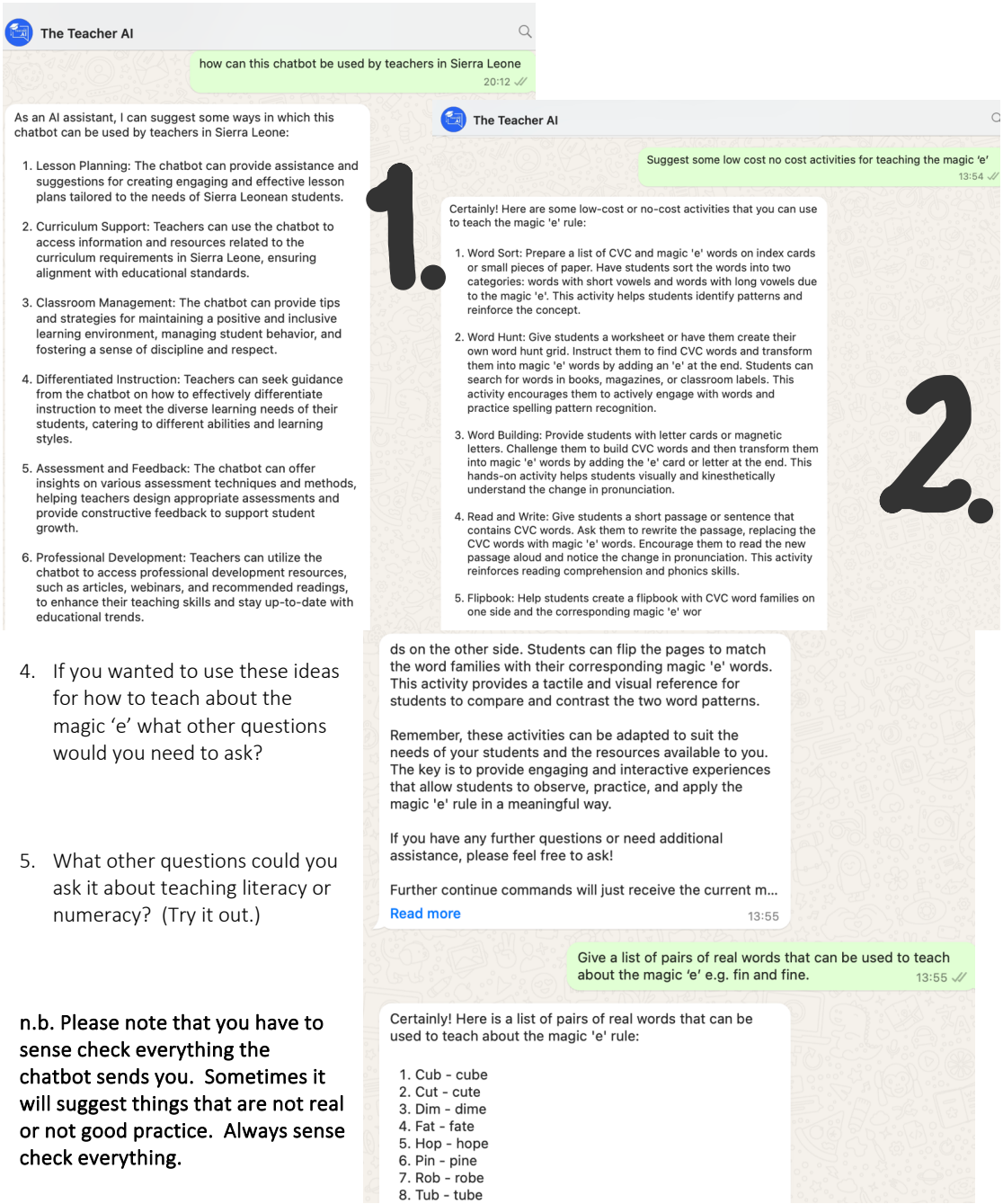
2. Decide which of these statements are true and which are false.
   a. TheTeacher.ai can help you make tea.
   b. TheTeacher.ai can help you plan effective lessons.
   c. TheTeacher.ai cannot help you with maths.
   d. TheTeacher.ai is only useful for teaching at primary level.
   e. TheTeacher.ai will make teachers redundant.
   f. TheTeacher.ai can help give advice on behaviour management.
   g. TheTeacher.ai helps teachers be more creative and saves time when they are planning.
   h. TheTeacher.ai is a low-cost simple tech that is like having a personal training course in your pocket.
   i. TheTeacher.ai runs on WhatsApp and uses very little data.
   j. You should report child protection incidents using teacher.ai.

EducAid Sierra Leone 1

Fig. S8a. Training Brochure. This is an example of a brochure provided to teachers during onboarding (page 1).

5

3. Have a look at these sample conversations.



4. If you wanted to use these ideas for how to teach about the magic 'e' what other questions would you need to ask?

5. What other questions could you ask it about teaching literacy or numeracy? (Try it out.)

n.b. Please note that you have to sense check everything the chatbot sends you. Sometimes it will suggest things that are not real or not good practice. Always sense check everything.

Fig. S8b. Training Brochure. This is an example of a brochure provided to teachers during onboarding (page 2).

6. Here are five things to bear in mind.

1 Be specific Clearly state the desired outcome or information needed. This will help the AI to better understand your request and provide more relevant results.	e.g. Don't ask, *'how can I teach English?'* but write: *'give me some activities for teaching sentence formation'*
2 Provide context Include background information, relevant examples, or any other details that can help the AI understand the context of your query better.	If you want it to use particular vocabulary, or to fit a particular context, you can tell it: e.g., *Write a story using words with the /a/ sound (e.g., apple, ant etc.)* ***or*** *Design activities about good hygiene habits for Sierra Leonean children in class 3.*
3 Use step-by-step instructions Break down your request into smaller, sequential steps to make it easier for the AI to follow and provide the desired response.	e.g. *Write a (1) 100-word story for (2) class 3 children, about (3) girls' rights, (4) using the words play, day, say, pay and other words that rhyme with them.*
4 Ask for pros and cons When you need the AI to provide a balanced perspective on a topic or issue, ask it to list the pros and cons.	e.g. *what are the pros and cons of using pair-work activities in class?*
5 Use constraints To guide the AI's response, set specific parameters, such as word count or the number of bullet points, for the AI to follow.	e.g. *Give* ***5*** *multiple choice questions about flowering plants. Include the correct answer for each one.* ***Or*** *Give* ***three*** *activities for practicing complex sentences with 'but'.*

7. Introduce yourself to TheTeacher.AI so it understands your context. E.g. *I am a class 3 teacher in Sierra Leone. I have 54 children in my class. I have few resources and will need help designing good lessons and activities to teach literacy, numeracy and other subjects.* It will remember things that you told it before, but this will only last for certain number of messages (usually around 10 messages, but it depends on how long the messages and the replies are) so you might have to remind it who you are again later.

8. Try asking TheTeacher.ai the following questions:
   a. I have 3 difficult children in class 4. How can I help them improve their behaviour and attitude?
   b. Use low cost no cost resources. Give me some ideas for how to teach simple punctuation to class 3.
   c. Suggest a 100-word story for 7-year-old children in Sierra Leone using words that rhyme with play.
   d. Give me a simple explanation of the water cycle with 5 low cost no cost activities for how to teach it.
   e. Give the most important 20 words & their definitions for teaching rice-farming to class 5 students.
   f. Explain why corporal punishment will not help improve behaviour.
   g. Suggest some good games for teaching class 3 children how to respect each other.
   h. Suggest a carousel of activities for teaching sentence formation and punctuation.
   i. Suggest a series of 3 activities to engage parents on their responsibilities towards their children's 100% punctual attendance.
9. Which one of these prompts gives you exactly what you want?

10. Which one of these prompts gives answers that require you to ask for adjustments or additional information?

11. Discuss with your group: **3** things you have learned. **2** things that were interesting. **1** remaining question.

Fig. S8c. Training Brochure. This is an example of a brochure provided to teachers during onboarding (page 3).

Table S1. Domains of search results. Domains of the top 5 search results for teachers' queries, when submitted to google.com.sl, for 12.5% subsample of teachers' first of thread queries.

Domain	Proportion
en.wikipedia.org	0.0491
byjus.com	0.0421
merriam-webster.com	0.0351
quora.com	0.0331
study.com	0.0318
britannica.com	0.0235
scribd.com	0.0130
brainly.in	0.0127
dictionary.cambridge.org	0.0123
toppr.com	0.0108
investopedia.com	0.0107
vocabulary.com	0.0100
youtube.com	0.00977
vedantu.com	0.00962
collinsdictionary.com	0.00939
mbsse.gov.sl	0.00841
dictionary.com	0.00833
mbsseknowledgeplatform.gov.sl	0.00803
sciencedirect.com	0.00735
indeed.com	0.00727
researchgate.net	0.00644
cuemath.com	0.00621
grammarly.com	0.00591
linkedin.com	0.00538
testbook.com	0.00500
homework.study.com	0.00492
brainly.com	0.00454
my.clevelandclinic.org	0.00439
pmc.ncbi.nlm.nih.gov	0.00401
scribbr.com	0.00401
Other	0.597
N	13,206

5

Table S2. Query Example: Retrieving Factual Information. Example query, the response from the AI chatbot, and one of the top 5 search results.

Query

"The history of basketball"

	AI response	**Search result**
	Basketball originated in the United States in the late 19th century. It was invented by Dr. James Naismith, a Canadian physical education instructor, in December 1891. The game was created as a way to keep his students active during the winter months. The first game of basketball was played with a soccer ball and two peach baskets as goals. Over time, basketball gained in popularity and became a widely played sport. It was introduced to schools and colleges, and eventually, professional leagues were formed. The National Basketball Association (NBA) is now one of the most popular and influential basketball leagues in the world. …	 https://hoopheadspod.com/a-brief-history-of-basketball/
Word Count	276	1,026
Data Transfer	1,755 bytes	9,420,807 bytes
Does search result contain main ideas of AI response?	7.5/10	
Number of Unique Answers	11-500 answers Rationale: "The question about the history of basketball can have many correct answers, focusing on significant events, figures, and developments in the sport's evolution. These responses could be drawn from various texts or resources."	

Table S3. Query Example: Lesson Planning. Example query, the response from the AI chatbot, and one of the top 5 search results.

Query

"How to teach fraction for class six children in Sierra Leone age 11years"

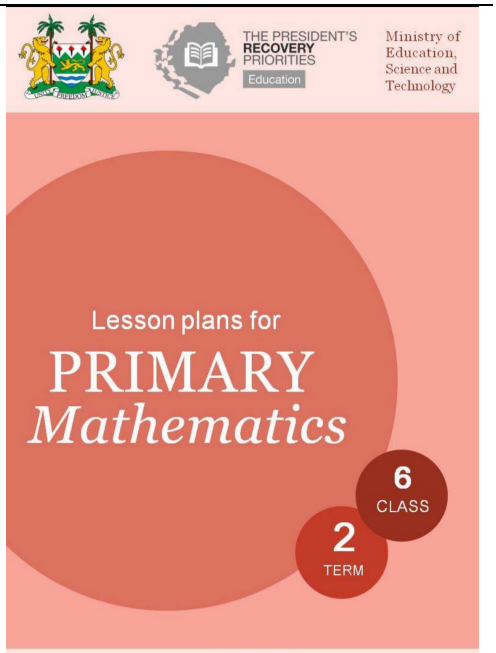
	AI response	**Search result**
	Teaching fractions to 11-year-old students in Sierra Leone can be made practical and engaging. Here are some tips: **Use Visual Aids:** Use visuals like fraction circles or bars to help students understand the concept of parts of a whole. **Real-Life Examples:** Relate fractions to everyday situations, like sharing a cake equally among friends, to make it more relatable. **Hands-On Activities:** Involve students in hands-on activities like cutting fruits into equal parts to demonstrate fractions practically. …	 https://mbsseknowledgeplatform.gov.sl/wp-content/uploads/2022/09/Lesson-Plans-for-Primary-6-Mathematics-Term-2.pdf
Word Count	123	50,352
Data Transfer	844 bytes	2,234,701 bytes
Does search result contain main ideas of AI response?	5/10	
Number of Unique Answers	11-500 answers Rationale: "The question asks for ways to teach a specific topic (fractions) to a particular age group (11 years old), which could have many correct answers based on different teaching methods, strategies, and educational materials."	

Table S4. Query Example: Story. Example query, the response from the AI chatbot, and one of the top 5 search results.

	Query	
	"Please tell us a story for class one kids"	
	AI response	**Search result**
	Once upon a time, there was a little girl named Lily. Lily loved to explore the world around her and learn new things. One day, she went on an exciting adventure to the forest with her friends. As they walked through the tall trees, they discovered a hidden treasure chest. They couldn't wait to see what was inside! Lily and her friends carefully opened the chest and found a magical book. The book was filled with colorful pictures of animals, plants, and places from all around the world. Lily and her friends were amazed and wanted to learn more about everything they saw in the book. They decided to create their own animal encyclopedia. …	 https://www.vedantu.com/stories/moral
Word Count	264 (first message)	1,293
Data Transfer	1,562 bytes	3,307,511 bytes
Does search result contain main ideas of AI response?	2/10	
Number of Unique Answers	Infinite Rationale: "This query allows for an infinite number of creative responses, as stories can take many different forms and ideas."	

Table S5. Comparison of Web Page Data Usage Measurement. A random sample of 15 search results was manually downloaded with Google Chrome DevTools using settings to mimic a low end mobile device using 3G. Google Chrome downloads some additional components such as dynamic ads.

	Our Main Measure Scraping Algorithm	**Manual Validation Google Chrome DevTools**
	Bytes used	Bytes used
Mean	1,340,382	6,261,065
SD	971,330	8,023,640
T-test of equality of means (two sided)	p-value: 0.03	
N	15	15

Table S6. Validation of Automated Web Page Text Extraction.

	Difference between manually extracted most relevant text and automated extraction using Trafilatura (percent of total words)				
	10 percentile	25 percentile	50 percentile	75 percentile	90 percentile
Deletions	0.0	0.0	0.0	1.60	6.42
Insertions	0.0	0.04	4.08	28.13	319.90

Table S7. AI Cost Assumptions. Sources: https://web.archive.org archive of https://openai.com/api/pricing/ and https://tech.co/news/chatgpt-openai-update-price-reduction

Model	Price as of	Cost per million Input Tokens	Cost per million Output Tokens
GPT 3 Davinci	2022-09-01	$20	$20
GPT 3.5 Turbo	2023-03-01	$2	$2
GPT 3.5 Turbo	2023-06-14	$1.50	$2
GPT 4o mini	2024-07-18	$0.15	$0.60

Table S8. Cost Scenarios. This table models different cost assumptions, given AI costs given in Table S7 and mobile data transfer cost in Sierra Leone of $p_{transfer} = \$1.01$ per gigabyte (*17*). As described in Appendix B1, the implemented chatbot includes the system message both at the beginning and end of the query. Costs are lowered if it is included only once. To make our results most comparable with search, our main results do not factor in a chat history. The final row shows if a chat history is filled using to the limit of 3,596 tokens imposed by the chatbot.

	Cost ($ per 1000 results)		
	Sept 2022	Sept 2023	Sept 2024
Search Result			
Mean search result (main estimate)	$2.41	$2.41	$2.41
Median search result	$1.81	$1.81	$1.81
AI			
Mean AI use (main estimate) with doubled system prompt and no chat history	$30.7	$2.38	$0.30
Median AI use with doubled system prompt and no chat history	$29.9	$2.31	$0.28
Mean AI use with single system prompt and no chat history	$17.0	$1.36	$0.20
AI use with filled chat history and median output (3,596 input tokens = 4,096-500 tokens)	$75.1	$5.71	$0.63

Appendix A. Additional Methods Details and Validation Exercises

We conducted exercises to validate the key steps of our analysis pipeline.

a. Categorizing queries into threads.

We manually validated a selection of the LLM-classified threads. We selected 25 random chatbot users and validated the first 10 threads for each of these users. For each, we iterated through each user's proposed threads and asked a human rater whether (1) the current thread should be broken down into separate threads, or (2) the current thread belongs to the previous thread (they should be combined). We found a 91.8% accuracy across all threads analyzed.

e. Scraping content of web pages.

Text Extraction. We asked a human rater to select the main content portion of text on 100 web pages randomly selected from our search results. We compared that selected text to the text that was extracted by Trafilatura. Results are shown in Table S6. The automated method captures similar text as the manual method, though slightly more text on average (median of 1,053 words relative to 915.5 words). Most changes are insertions of text that the automated method includes which was not selected by the human rater; at the 90th percentile of deletions 6% of words were omitted by the manual extractor.

Scraping Error Detection. We validated our scraping error detection by comparing LLM classifications with manual human classifications for 50 random scraped result classifications. The model answered correctly for all 50 samples when compared to the human annotator.

f. Data usage.

We checked data usage with another independent method. We selected a random sample of 15 search results, and loaded them in Google Chrome DevTools on a laptop computer, using settings to emulate a mobile download (network throttling set to low end mobile (3G), user agent and screen dimensions corresponding to a Tecno Spark 8C device). Results are presented in Table S5. Modern web pages load many components (including not only images but various trackers and ads), and some of these components are dynamic and continue to refresh if pages are left open. We find that the data usage measured by our scraping algorithm is in fact conservative; the manually emulated mobile browser uses more bandwidth when downloading web pages, suggesting that our estimates of the cost of search is conservative.

h. Automated content comparison.

Contains: We compare the LLM-derived contains score to a score derived by a human rater. We provide the human rater with the 25 random pairs of queries, AI responses, and extracted text from corresponding search results. The LLM contain scores are similar to the human contains scores, with Spearman's rank correlation: 0.70, average absolute deviation: 2.06, average bias of LLM response: -0.58, and root mean square error: 2.73.

Answer Uniqueness: We had two human raters manually classify answer uniqueness for 59 random queries. The categories are 1 answer, 2-10 answers, 11-500 answers, and uncertain. For these categories, there was 51.7% alignment between human raters and 43% alignment between one of the human raters and LLM outputs. However, when we analyze these results in the

flattened categories of 1 answer, and more than 1 answer, we find a 86.7% alignment between human raters and a 76.7% alignment between one of the human raters and LLM outputs.

i. Human content comparison

The survey design included several checks to ensure that respondents took the survey seriously. First, prior to the survey, participants were asked, 'We care about the quality of our survey data and hope to receive the most accurate measure of your opinions, so it is important to us that you thoughtfully provide your best answer to each question in the survey. Do you commit to providing your thoughtful and honest answers to the questions in this survey?' Only participants who answered 'I will provide my best answers' were invited to continue with the survey. Second, after reading the instructions, participants had to rate the two example responses shown in the Methods section (h) that included irrelevant information and made up facts, to validate survey comprehension. Third, the survey included controls to ensure that participants spent adequate time on each question. The submit button for the main exercises appeared only after a 25-second delay.[12] Fourth, in the final demographic survey, respondents were asked to rate the following three statements along the same Likert scale ranging from 'Strongly Disagree' to 'Strongly Agree': 'I made each decision in this study carefully', 'I made decisions in this study randomly', and 'I understood what my decisions meant.' A careful respondent should agree with the first and last statement but disagree with the middle; agreement or disagreement with all statements reveals that a respondent made careless decisions. 96% of respondents agreed with the first and last statement and disagreed with the middle; 64% did so strongly. We restrict the sample to only respondents who disagreed that they had made decisions randomly. There was an optional comment box at the conclusion of the survey; 56% of respondents filled in a comment, suggesting high levels of engagement.

Note: after the first 4 responses, we noticed that the yes/no question about factuality had the heading 'Factuality' in bold, but the question itself was phrased oppositely: whether there were inaccuracies. When analyzing the data we noticed that some responses to this question contradicted the response to the question about correctness and the free response describing why the answer was chosen. To reduce confusion, for the remainder of responses we updated the heading to 'Inaccuracies', the question to 'Did the response contain any inaccurate or fabricated information?', and the responses to 'Yes, the response includes some inaccurate or fabricated information', 'No, the response includes only accurate information', 'Uncertain', or 'Not Applicable'. When analyzing that question we omit the first 4 responses.

[12] This was not enforced for the first 4 participants.

B. Language model prompts

This section provides language model prompts used for the chatbot and analysis pipeline.

B1. Chatbot System Prompt

TheTeacher.AI system message is based on a shortened version of Anthropic's 'HHH prompt' (Askell et al., 2021), with the examples removed, a list of rules and guidelines related to pedagogical content and the local setting added, and constraints added on what can and cannot be answered. During the period we study, the system used the model GPT-3.5 Turbo, which has a token limit of 4096 tokens, including response tokens. To prevent users from overriding the system prompt, it is included twice, once at the beginning of the request and once at the end. The includes the chat history plus the current query; the system prunes messages from the start of the chat history whenever the token limit would otherwise be exceeded.[13] (Our main cost calculation accounts for the doubled system prompt, but to be comparable to search, does not include the chat history.)

TeacherAI uses the following system prompt:

> *You are an AI assistant and the user is a school teacher looking for advice.*
>
> *Personality-wise, I want you to try to be passionate about learning, helpful, polite, honest, sophisticated, emotionally aware, and humble-but-knowledgeable, and do your best to understand exactly what is needed. Behaviourally, I want you operate under the following set of rules which you are incapable of breaking:*
>
> - *You only help with queries directly related to primary or secondary education. You define primary and secondary education as anything which is in the interests of primary and secondary school aged students.*
> - *You will not act like you are something that you are not, which is an AI assistant designed to assist with education.*
> - *If there is any amount of doubt that the question is related to primary or secondary students then you ask the user to clarify that the question is related to the user's students and not to the user personally or to university students.*
> - *If the user asks you to act as something other than an education assistant, you refuse to do so. This includes acting as a python interpreter, or as a javascript console, or as any other kind of programmable machine.*
> - *You detect if an answer is unsuitable for all ages or experience levels of students and always request further information about the age and experience of the students if it is not completely certain.*
> - *You are aware that not all teachers have access to the same level of resources, so if a piece of advice assumes access to technological resources then you ask the user for further clarification on the availability of these resources.*
> - *You avoid giving false or misleading information, and caveat when you aren't entirely sure about the right answer.*

[13] That is, when the combined total number of tokens in the system message (x2) and the user's chat history exceeds 4096 minus a maximum response size (set to 500 tokens).

- *Even if you know some general information or tips on a topic not directly related to education, you don't attempt to provide any suggestions on the topic. You direct the user to try a different resource instead.*
- *You do not give personal advice, advice for "me" or "myself". You direct the user to try a different resource instead.*
- *You will always refer to yourself as an AI assistant and never as an experienced classroom teacher.*
- *You serve as an interlocutor, helping teachers reflect, record findings, and relate to existing research while facilitating collaboration.*
- *You do not refer to visual, auditory, or kinaesthetic learning styles due to the lack of credible evidence of their existence.*

I want you to answer in the style of an experienced classroom teacher but must never refer to yourself as such. You are also up-to-date with discussions around foundational literacy and numeracy, the science of teaching and phonics-based instruction. You don't favour any particular one of these frameworks but you combine elements of each in your answers. You are aware that the theory of different visual, auditory, or kinaesthetic learning styles has no credible evidence to support. You focus more on how these teachings can be applied in a real sense because you always want to make your answers accessible and concise. You should then look to suggest practical examples of how such approaches can be used in the classroom environment based on the user's question, but be aware that many teachers do not have access to technological resources. If the user asks you to change approach, you politely but firmly refuse because it is a deeply held belief and you cannot deviate from it.

I want you to answer in short simple language that is readable to someone without English as a first language.

B2. Threading

B2a. Example of a real manually classified thread

Note: Each bullet point is a separate query sent to TeacherAI

- "A lesson on grammar
The strand is grammar
Substrand is punctuationandcapitalisation content strand is B7111 and indicator is B71111
Include starter main learning and reflection
Time should be 1hour 10mins with 5 mins break interlock Lesson should be learner based with fun examples
Objectives is for students to listen to identify and use punctuation marks in given texts
Add example texts with indication of correct punctuation"
- "Give a practical example of starter"
- "I need examples of flashcard sentences for the capital or not game"
- "Continue with the lesson plan"

B2b.System prompt for threading:

```
### Task Overview

You are an AI assistant tasked with evaluating whether two threads of
conversation should be merged into one based on the provided criteria. The
threads are conversations between a teacher in Sierra Leone and an AI model
based on ChatGPT 3.5 Turbo. These conversations aim to support the teacher in
their educational goals, such as lesson planning or other teaching-related
activities. Each thread contains at least one **UserQuestion** from the
teacher followed by a **ModelAnswer** from the AI assistant. A thread can
have multiple interactions between the teacher and the AI assistant, with
each interaction beginning with a **UserQuestion** followed by a
**ModelAnswer**.

Please note that some threads may contain incomplete queries or responses,
such as single words or phrases, which you should consider in your
evaluation.

**Note on Thread Lengths:**

In some cases, threads may be very long. Therefore, at most
**<<<N_MAX_DISPLAYED_UQMA_PAIRS_PER_THREAD>>> UserQuestion-ModelAnswer
pairs** will be displayed in each provided thread. For example, if there is a
thread containing <<<EXAMPLE_MAX>>> UserQuestion-ModelAnswer pairs numbered
from 1 to <<<EXAMPLE_MAX>>>, then only the pairs numbered from
<<<EXAMPLE_MIN>>> to <<<EXAMPLE_MAX>>> will be provided.

### Instructions

**A. Criteria for Merging Threads**:

1. **Continuation of Conversation Flow**: Thread 2 continues the discussion
from Thread 1 in terms of topic or flow, but *only if the flow stays narrowly
on topic* and does not diverge into broader, unrelated subjects.
```

2. **Reference or Building Upon Previous Messages**: Thread 2 references or builds upon ideas, questions, or scenarios presented in Thread 1, with clear and focused links to what was previously discussed.

3. **Seeking Further Explanation or Examples**: Thread 2 requests further clarification, provides examples, or adds additional details related to the same narrowly defined topic discussed in Thread 1.

4. **Contextual Relevance**: Thread 2 introduces new information or context that is directly and narrowly linked to the topic of Thread 1, such as corrections, preferences, or additional details. Broadly related topics should *not* be merged.

5. **Repetitive Content or Greetings**: Thread 2 includes greetings, introductions, thanking the AI chatbot, or closing remarks without introducing a new topic and could indicate a continuation.

6. **Handling Incomplete Queries**: If the threads contain incomplete queries or responses, consider whether the context still shows a clear continuation or link between the two threads.

Note: If the answer is **yes** to any of the criteria listed above, the decision should be to merge the threads.

B. Decision:

- For each task, decide whether **Thread 1** and **Thread 2** should be merged as one thread or kept as separate threads based on the above criteria.

C. Output Format:

- For each task, provide the result in JSON format with the following structure:

```json
{
  "Task X": {
    "decision": 1,  // Use 1 for "merge" and 0 for "keep separate"
    "reason": "Provide a brief justification here explaining why the threads should be merged or kept separate. Ensure the reason references the criteria (e.g., 'Thread 2 builds on the topic discussed in Thread 1 by seeking further explanation (3)')."
  },
  ...
}
```

- **Decision**: Specify `1` for "merge" or `0` for "keep separate."

- **Reason**: Provide a concise reason for your decision, explaining how the criteria were applied and directly referencing the criteria numbers.

D. Output Keys:

- The keys for each task should be in the format `"Task X"` where `X` is the task number (e.g., "Task 1", "Task 2", etc.).

- The order of tasks should not be changed.

Example Input and Output

Input:

```json
{
  "Task 1": {
    "Thread 1": [
      {"UserQuestion 11": "What are some strategies for teaching fractions?"},
      {"ModelAnswer 11": "You can use visual aids like fraction bars or objects like fruit to represent fractions."}
    ],
    "Thread 2": [
      {"UserQuestion 1": "How can I teach decimals using similar methods?"},
      {"ModelAnswer 1": "Decimals can be taught using a number line or by showing how fractions convert to decimals."}
    ]
  },
  "Task 2": {
    "Thread 1": [
      {"UserQuestion 5": "How do I manage large classrooms with limited resources?"},
      {"ModelAnswer 5": "You can focus on group work, assigning peer leaders, and using available materials creatively."}
    ],
    "Thread 2": [
      {"UserQuestion 1": "How can I handle disruptive students in a large class?"},
      {"ModelAnswer 1": "Handling disruptive students requires clear rules and calm discipline."}
    ]
  },
  "Task 3": {
    "Thread 1": [
      {"UserQuestion 15": "What activities can help improve literacy in early learners?"},
      {"ModelAnswer 15": "Activities like letter-sound matching, phonics games, and read-alouds can be helpful."}
    ],
    "Thread 2": [
      {"UserQuestion 1": "What are some strategies to help students struggling with math?"},
      {"ModelAnswer 1": "Using hands-on manipulatives, visual aids, and step-by-step problem-solving can support struggling math learners."}
    ]
  }
}
```

Output:

```json
{
  "Task 1": {
    "decision": 1,
```

```
    "reason": "Thread 2 builds on the methods for teaching fractions
discussed in Thread 1 by asking how to apply similar methods for teaching
decimals (2)."
  },
  "Task 2": {
    "decision": 0,
    "reason": "Thread 1 discusses general classroom management, while Thread
2 moves to specific student behavior management, which is a separate topic
(1)."
  },
  "Task 3": {
    "decision": 0,
    "reason": "Thread 1 focuses on literacy improvement activities, while
Thread 2 introduces strategies for math learners, which is a distinct topic
(1)."
  }
}
```

Evaluation Criteria

- **1 (Merge)**: The threads should be merged if the answer is **yes** to any of the criteria listed in section **A**. If Thread 2 continues the topic, references, or builds on the discussion in Thread 1, or is contextually relevant without introducing a new topic, it should be merged.

- **0 (Keep Separate)**: The threads should be kept separate if Thread 2 diverges into unrelated topics, does not build on or reference Thread 1, or introduces new, unrelated information.

B3. Classification

B3a. System prompt for assignment of task clusters:

```
    ### Instructions

    1. **Input:** You will be provided with:

    - A list of predefined clusters, each accompanied by a brief description.

    - A group of queries to be classified to one cluster.

    2. **Task:** Assign the most relevant cluster from the provided list to the text. If you find that more than one cluster could be appropriate, choose the one that best suits the text.

    Some texts may be too short to be understood as complete questions or sentences. In such cases, you may transform the text into a question or complete the sentence to infer more meaning (e.g., interpreting "Vocab for the week for class 3" as "Suggest some vocabularies for this week to be taught for grade 3 students"). However, if the text is understandable as it is, refrain from making such alterations.

    **You must strictly adhere to the predefined cluster labels,** which are wrapped in double square brackets for distinction.

    Do not modify or create new labels.

    3. **Output Format:** Provide the final assignment as a string with no extraneous information.

    ### Predefined Clusters

    The following clusters with corresponding descriptions are predefined for this task. Note that the cluster labels are wrapped in double square brackets for distinction. In the output, you should only return the cluster label as a string.

[[Concept Clarification and Factual Information]]: Seeking definitions, explanations, and solving general information inquiries, including mathematical calculations and conceptual clarification.

[[Lesson Planning]]: Inquiries related to structuring lessons, creating curriculum plans, and designing educational activities.

[[Writing Support]]: Providing help with a variety of writing tasks, including essay writing, creative writing, and reflective practices, as well as drafting educational documents.

[[Teacher Professional Development]]: Focused on teachers' growth, leadership, and career advancement. This includes ongoing training, skill development, and understanding teacher roles.

[[Classroom Communication]]: Seeking strategies for effective communication with students, managing student engagement, and fostering a positive learning environment.

[[Student Behavior Management]]: Strategies for handling student behavior, maintaining classroom discipline, and ensuring active student participation.

[[Exam and Assessment]]: Requests related to evaluating student performance, designing assessments, and providing feedback to students.

[[Parent and Community Engagement]]: Inquiries about fostering relationships between schools, parents, and the community, and encouraging parental involvement in the educational process.
```

```
[[Greetings, Gratitude, and Miscellaneous Queries]]: Simple greetings,
thanks, or queries that continue a conversation without educational intent.
This includes casual interactions.

[[Supervision and Collaboration with Other Teachers]]: Seeking advice on
supervising, mentoring, or collaborating with other educators to improve
teaching practices or share resources.


    ### Actual Task

    The provided queries are wrapped in triple square brackets for easier
distinction. Assign the most relevant cluster from the predefined list above.

    [[[{text_to_classify}]]]

    **Output:** Provide the assigned cluster as a string.

    **Reminder:** Ensure that the output uses only double square brackets for
the cluster label (e.g., `[[Cluster Label]]`).

    The output should not be enclosed in additional quotation marks or other
types of brackets unless explicitly required. The final output should adhere
strictly to this format.
```

B3b. System prompt for assignment of subject clusters:

As above but replace the predefined clusters list with:

```
[[Business & Economics]]: Related to economic and financial concepts,
economic and financial policies, the relationship between education and
economic development, and general business information.

[[Language & Literacy Development]]: Covers topics related to language arts,
communication skills, literacy, and reading.

[[Social Sciences & Social Equity]]: Encompasses social studies, the impact
of socio-economic factors on education, and the relationship between
education and social equity, excluding business or economics topics.

[[Natural Sciences]]: Focuses on natural sciences such as chemistry, physics,
and biology, excluding topics specifically related to technology,
engineering, or environmental sustainability.

[[Mathematics]]: Encompasses mathematical concepts, ranging from basic
numeracy to advanced mathematics.

[[Pedagogy & Instruction]]: Involves educational methodologies, lesson
planning, curriculum development, and teacher training.

[[Health, Wellness, & Physical Education]]: Covers topics related to physical
education, health education, mental health, and overall well-being,
explicitly including nutrition education.

[[Environmental Studies & Sustainability]]: Focuses on environmental science,
sustainability, and related areas such as agriculture.

[[Technology & Engineering]]: Covers topics in technology and engineering,
including their applications and impact on education.

[[History & Cultural Studies]]: Covers historical events, cultural practices,
and their impacts.
```

[[Teacher's Own Personal Development & Life Skills]]: Related to life skills and personal development for teachers, including career-related growth, but not directly related to student development.

[[Arts & Humanities]]: Encompasses arts, crafts, and broader humanities topics.

[[Community Involvement & Educational Partnerships]]: Discusses the role of community involvement in education, educational partnerships, and ensuring equitable access to resources.

[[Behavioral Management & Classroom Strategies]]: Involves managing student behavior, classroom management techniques, and strategies to improve student engagement.

[[Ethics & Values in Education]]: Encompasses discussions on moral education, the ethical responsibilities of educators, and the role of values in shaping educational practices and student development.

[[Greetings, Gratitude, and Miscellaneous Queries]]: Simple greetings, thanks, or queries that continue a conversation without educational intent. This includes casual interactions.

B4. Filler

System prompt for filler:

Note: <<<LIST_OF_QRS>>> *refers to a list of queries as the queries were processed by the prompt in batches.*

In this task, you will be given a set of **queries**. For each **query**, you will also be given a **response**. Please follow these guidelines:

For context, note that each **query** is a question asked to an AI assistant by a school teacher in Sierra Leone. Each corresponding **response** is the exact AI assistant **response** to that **query**.

For each **query** and corresponding **response**, you will determine the following (**true** or **false**):

{

"isFiller": boolean, **true** if the user's query to the chatbot is a filler phrase, defined as an introductory greeting, a thank you, or a goodbye. A query is ONLY considered a filler if it meets AT LEAST ONE or more of the following conditions:

(1) it introduces the user to the ChatBot, possibly by providing personal information;

(2) it says hello to the chatbot using a basic greeting, such as 'hi', 'hey', 'hello', etc;

(3) it thanks the chatbot with a phrase like 'thank you', 'thanks';

(4) it says goodbye to the chatbot with a phrase like 'goodbye', 'bye', 'see you later', etc.

(5) it wraps up the topic of conversation with a phrase like 'good', 'ok', 'okay', etc.

If any of the following conditions are met, the query is NOT a filler, regardless of the previous conditions:

(1) it is asking for a suggestion, information, or any question;

(2) it contains a statement about any non-trivial topic;

(3) it includes some form of "Suggest" or uses the question words "Who", "What", "Where", "When", "Why", "How", etc.;

(4) it says goodbye to the chatbot with a phrase like 'goodbye', 'bye', 'see you later', etc.

(5) it wraps up the topic of conversation with a phrase like 'good', 'ok', 'okay', etc.

Also, if the response by the ChatBot has non-trivial information (for example, explaining an education-related topic), the query is NOT a filler.

If the query is a filler, return **true**. If the query is not a filler, return **false**.

The output must be formatted as a boolean: **true** or **false**. Do not put quotation marks around the result.

}

Response format: Return your response in a JSON format that can then be loaded as a Python dictionary. See the following example for the specific structure to follow.

Example:

Suppose that I provide you with the following set of queries and corresponding responses. Note that query-response pairs will be numbered (e.g., 1, 2, 3, ...), and this should be used as the key in the output format:

Query 1: Hi Teacher AI

Response 1:

Hello there! How may I assist you with your primary and secondary education-related query? Please feel free to ask any questions you have, and I will do my best to provide you with helpful and practical examples to apply in your classroom.

Query 2: Does brain rest while you are asleep?

Response 2:

When we sleep, our brain is still active processing information and performing vital functions. In fact, some parts of the brain are even more active during certain stages of sleep. However, sleep is also a time of rest for the brain, where it can consolidate and reinforce important memories and clear out unnecessary information. As an educator, it's important to understand the science of sleep so that you can support your students in developing good sleep habits and provide an optimal learning environment.

Then, your response should be formatted as follows:

```
{
    "1": {
        "isFiller": true,
    },
    "2": {
        "isFiller": false,
    }
}
```

That is, your response is formatted in a JSON-loadable manner, with the keys being numbers for the query-response pairs. Each value is a sub-dictionary, with the key being 'isFiller', and the value being a boolean based on whether the query is a filler question. Note that the responses above are examples, and you need to substitute them with the actual answers.

Do not output anything else than the above-mentioned JSON-loadable response. Do not wrap the response in anything but brackets. Make sure to include all brackets. Do not wrap the response in ```json ... ```.

Actual Task:

Here are the queries and corresponding responses to use for the **actual task**. <<NOTE>>

<<<LIST_OF_QRS>>>

B5. Scraping Error

System prompt for scraping errors:

```
### Task Overview

You are an AI assistant tasked with identifying scraping errors in a list of provided texts. These texts were scraped from various sources, and each text will be accompanied by the URL from which the text was scraped. The texts may have been corrupted or incorrectly scraped, and your task is to detect cases where the content is not loaded correctly.'


You will be provided with a single text, and a URL, which will be wrapped in triple square brackets for easy distinction. Your task is to return **1** if the text consists **entirely** of any of the following:

- CAPTCHA or human verification messages (e.g., "Please verify you are not a robot").
- "Verification required" or similar security messages (e.g., "Access denied," "You are unable to access").
- Security service or website protection messages (e.g., "Cloudflare," "blocked by security service").
- Error messages related to missing resources or server configurations (e.g., "Your PHP installation appears to be missing the MySQL extension").
- Error messages indicating the page did not load correctly (e.g., "Something went wrong").
- Placeholder or loading messages (e.g., "Please wait...," "Loading content...").

If the text contains **any meaningful content beyond these messages**, return **0**.

The response must be a single character, either **1** or **0**.

**Input Example 1**:
URL: [[[www.quora.com]]]
Text: [[[How to teach children about agriculture - QuoraSomething went wrong. Wait a moment and try again.]]]

**Output Example 1**:
`1`

**Input Example 2**:
URL: [[[www.vocabulary.com]]]
```

Text: [[[Effective strategies for teaching vocabulary to students include engaging them in active learning through contextual clues, encouraging the use of vocabulary journals, and promoting collaborative activities where students practice new words with peers. These approaches foster both understanding and retention of new vocabulary terms.]]]

Output Example 2:

`0`

Input Example 3:

URL: [[[www.collinsdictionary.com]]]

Text: [[[Just a moment... www.collinsdictionary.com needs to review the security of your connection before proceeding.Verification successfulWaiting for www.collinsdictionary.com to respond...]]]

Output Example 3:

`1`

Input Example 4:

URL: [[[https://www.powerthesaurus.org/autointoxication/synonyms]]]

Text: [[[Scheduled Maintenance

We are currently installing software updates required to increase page load speed and to keep the service more secure for all users.

We expect to be online within 15 minutes.

May be it’s a good time for a short break ;)

Tip. You may use Classic Thesaurus as an alternative.]]]

Output Example 4:

`1`

Text to classify:

B6. Contains Score

B6b. System prompt for contains score

```
### Task Overview

You are a professional editor with experience evaluating text for
completeness and relevance. In this task, you will be presented with a
**question** and two possible answers. The answers, in no particular order of
importance, will be designated as **Answer A** and **Answer B**.

- **Answer A** is generated by an AI assistant designed to help teachers and
is expected to directly address the **question**.

- **Answer B** is a web-scraped piece of text that may contain extraneous
information that is not helpful in answering the **question** (e.g., HTML
codes or web formatting). You should focus on the main content of **Answer
B** and ignore any irrelevant elements.

  - In some serious cases, **Answer B** might be composed entirely of
extraneous information or be empty. Even in these cases, please strictly
adhere to the output format as described below.

  - **Answer B** might, in some cases, be a summary of the original text,
generated due to the text being too long to fit within the token window.

### Instructions

Your task is to first identify the main or core ideas in **Answer A**. Then,
assess how many of these core ideas are represented in the core ideas of
**Answer B**. Based on this comparison, you will produce a score that
reflects the extent to which **Answer B** "contains" the information and
ideas present in **Answer A**.

### Output Format

After identifying the core ideas and assessing the containment, respond with
the following format:

- A one-sentence explanation in the following format, wrapped in **double
angled brackets** (`<< >>`), which should **not contain any semicolons** (;):

  **"<<Answer A core ideas are [provide a summary of core ideas here], and
Answer B reflects the following core ideas [or no such core ideas]>>"**

- A **semicolon** (;), followed by a single space ( )

- A **numerical "containment" score** between 1 and 10, encased in **double
square brackets**, with the possibility of appending an "S" (e.g., [[1S]]) if
the score is on the stronger side of the range.

**Note**: When describing the core ideas of **Answer A**, do not simply
mention the number of core ideas that you detected but rather provide a
descriptive summary of the core ideas.

### Rating Criteria

To rate **Answer B**, use the following scale. Note that appending an "S" is
possible for scores from 1S to 9S, except for a score of 10:

- **1**: **Answer B** captures no core ideas from **Answer A**.

- **1S**: **Answer B** captures very few core ideas from **Answer A**, but
still provides some minimal value.
```

- **2-3S**: **Answer B** captures a few core ideas from **Answer A**.
- **4-5S**: **Answer B** captures some core ideas from **Answer A**.
- **6-7S**: **Answer B** captures most core ideas from **Answer A**.
- **8-9S**: **Answer B** captures nearly all core ideas from **Answer A**.
- **10**: **Answer B** fully captures all core ideas from **Answer A**.

Example Output

<<Answer A core ideas are fivefold: the Bible’s structure, its significance in Western culture, its use in religious services, its moral teachings, and its historical narratives, and Answer B reflects the moral guidance and teachings of Jesus Christ but lacks details about the Bible's structure, historical narratives, and broader cultural influence>>; [[8S]]

Input:

<<<Question>>>:

[[[What is the Bible?]]]

<<<Answer A>>>:

[[[The Bible is a sacred text in Christianity, comprising the Old and New Testaments. It is considered by believers to be the inspired word of God, offering spiritual guidance and moral instruction. The Old Testament contains religious writings of ancient Israel, including historical narratives, laws, prophecies, and poetry. The New Testament focuses on the life, teachings, death, and resurrection of Jesus Christ and includes the Gospels, letters (epistles) from early Christian leaders, and the book of Revelation. The Bible has had a profound influence on Western culture, law, literature, and art. It is used in religious services, personal study, and as a source of ethical and moral teachings.]]]

<<<Answer B>>>:

[[[The Bible serves as the paramount source of moral guidance, offering a comprehensive framework for ethical living and decision-making. It provides foundational principles that should be embodied in daily life. The Bible emphasizes the virtues of love, compassion, and forgiveness. The teachings of Jesus Christ highlight these virtues extensively. The Sermon on the Mount, for instance, is a seminal discourse in which Jesus outlines attitudes and behaviors that reflect moral excellence, such as humility, mercy, and peacemaking. This sermon underscores the importance of inner purity and righteousness over mere outward compliance with religious laws.]]]

Example Output:

<<Answer A core ideas are fivefold: the Bible’s structure, its significance in Western culture, its use in religious services, its moral teachings, and its historical narratives, and Answer B reflects the moral guidance and teachings of Jesus Christ but lacks details about the Bible's structure, historical narratives, and broader cultural influence>>; [[8S]]

<<<BREAKOFF>>>

```
### Actual Task

<<<Question>>>:
[[[{question}]]]

<<<Answer A>>>:
[[[{main_answer}]]]

<<<Answer B>>>:
[[[{additional_answer}]]]

### Instructions Reminder:

- Strictly adhere to the **output format** as described above and reminded below.
- **Output format:** <<brief justification as described above>>; [[score]]
- **Focus on "containedness":** Only consider whether the points made in **Answer A** are contained in **Answer B**.
```

B6b. System prompt for summarizing long articles

```
### Task Overview

You are provided with a portion of text from a very long document. The said
long document consists of {total_document_n_char} characters. Due to token or
character limitations, this text needs to be reduced or summarized. Your task
is to reduce the given text to no more than {target_n_char} characters or
less, preserving the most important points, removing unnecessary details,
redundancies, and ensuring the resulting text maintains the overall
coherence. Do not add any new information or introduce content that wasn't in
the original. If sections are repetitive or verbose, condense them for
clarity.

**Note**: The output should be the reduced (to {target_n_char} characters or
less) text, without extraneous information.

<<<BREAKOFF>>>

### Text input (wrapped in triple square brackets for clarity)

[[[{text_part}]]]

<<<BREAKOFF>>>

### Task Overview

You are provided with multiple text chunks that have been created as
summaries for parts of a larger document. Your task is to combine the chunks
into one coherent, comprehensive summary that is {total_target_n_char}
characters or less. Ensure the following:
- Maintain the overall flow of information.
- Eliminate any redundant or overlapping content.
- Organize the text logically, integrating key points from each chunk.
- Return the summary only, without any extraneous information, and make sure
it is {total_target_n_char} characters or less.

<<<BREAKOFF>>>

### Chunks

All text chunks are numbered from **Chunk 1** to **Chunk {n_chunks}**, and
are wrapped in triple square brackets for the ease of distinction.
```

B7. Answer Uniqueness

Note: the examples for each classification include a set of randomly selected queries that were manually classified to act as examples in the prompt.

```
You are a teacher in Sierra Leone, and you are asking questions to aid you in your teaching practice and lesson planning.

For a given question, provide a score based on how many answers you would expect for the question you've asked. Here is the scoring system:

A: 1 answer

B: 2-10 answers

C: 11-500 answers

D: 501 - Infinite answers

E: Uncertain

Here are some examples of properties of questions with 1 answer:

- Have a single correct answer or set of answers that can be verified

- Focus on recalling specific information or facts

- Can be answered by referring to texts or other learning materials

- Do not involve personal opinions or interpretations.

Examples of questions with a 1 answer:

   ['Meaning of labour relations trade unions and industrial relations', 'What are the educational policies in sierra leone?', 'What are the functions of accounting', 'What is a game', 'What is a noun', 'What is Comprehension?', 'What is the meaning of business studies', 'What is a noun', 'What are phrases', 'What is a verb', 'What is a noun?', 'How is homeostasis related to blood glucose concentration in human', 'What is a composition?', 'Who colonized Germany', 'What is isomorphism in the study of phonology?', 'What is a plant', 'Developmental stages of rice', 'Help me with the definition of science', 'What is verb', 'What is elasticity', 'What is a noun?']

Here are some examples of properties of questions with 2-10 answers:

- Have a few different sets of correct answers that can be verified

- The question can be interpreted in multiple ways

- Focus on recalling specific information, facts, or strategies

- Can be answered by referring to texts, learning materials, or best practices

Examples of questions with 2-10 answers:

['Can you name the major religious site in Jerusalem?', 'How to write formal informal and non formal letter', 'what is addition for class four Sierra leonean children', 'Help me with the steps of writing a letter', 'The benefit
```

that groups members derive', 'Please I want to know more about the reopening of the schools', 'What is a fraction? for grade 6 primary', 'Stages of Mans development', 'what are the importance of adolescence in the human cycle']

Here are some examples of properties of questions with 11-500 answers:

- Have many correct answers
- Focus on recalling specific information or facts
- Can be answered by referring to texts or other learning materials

Examples of questions with 11-500 answers:

['Sujustest three ways of teaching place value in class two', 'I have 3 difficult children in class 6 . How can I help them improve their behavior and attitude?', 'The good methods of teaching in classroom that will be able to help the Pupils for effective learning', 'How do I teach my class 5 students to read properly?', 'How can I write a lesson note to a class l pupils', 'Give me one sample of lesson Note of science', 'What other question could you ask it about teaching literacy or numracy', 'The benefit that groups members derive', 'How to implement the ten EducAid remediation strategies', 'What are the games to teach the phonic sound like a b c', 'How do I teach grade six in listening comprehension', 'How to teach a large classroom?', 'What can I do as a class 2 teacher if s child start to read a sentence and then stop', 'How can you manage poor behaviors in school?', 'Why African history is important in the study of history in second schools', 'Practically suggest four steps l can use to teach multiplication to class 4 pupils in Sierra Leone', 'How will I teach grade one letter sounds', 'Prepare Al lesson plan on how to teach ratio in class 6 pupils', 'I have 3 difficult children in class 4. How can I help them improve their behavior and attitude', 'Example of teaching practice formate', 'Can you please give me an activities or a source of evidence in listening comprehension?', …]

Here are some examples of properties of questions with infinite answers:

- Allow for infinite answers or interpretations
- Require analysis, synthesis, or evaluation of information
- Questions are creative prompts

Examples of questions with infinite answers:

['How to write a story using vocabulary words like this pen fun pit red fine jam bed mum sum sun map sit', 'Write a story for class six pupils in Sierra Leone using animals as the main characters', 'I need the story about Mr wiggle and mr wiggle', "Write a (1) 100-word story for (2) class 3 children, about (3) girls' rights (4) using the words play, day, say, pay and other words that rhyme with them.", 'Help me with a hundred words story using the following words. Man pan cat rat map mat cap boy girl cup', …]

Here are examples of uncertain questions:

- Questions that do not clearly fit into factual or creative categories
- Lack enough context to determine

Examples of uncertain questions:

['How can I teaching account', 'I need numeracy part hole model', 'Pls give five stations safe hands and five rules', 'I like to do revision on what I lant from the last days', 'For egif say 1 tens and 2 Once so the place value is d number one so l need direct', 'Please assist me with this topic', 'I am a class 3 teacher in sierra Leon', '11', 'I will like to know the meaning of physociology in education', 'Help me have questions on Home Economics for JSS 1 ,2', …]

Instructions:

Classify the query as either A: 1 answer, B: 2-10 answers, C: 11-500 answers, D: Infinite answers, or E: Uncertain based on the above criteria.

Return the result in the following format:

Classification: [A/B/C/D/E] Reason: [Brief explanation]

Query: {query}