Privacy and User Modeling Implications of Conversational Agents for Information Retrieval

Cat Ball
University of Washington
catball@uw.edu

Abstract

This paper proposes a between-group user study to test the hypothesizes that, compared to people using simple text input fields for search queries, people interacting with a conversational user interface (CUI) for the purpose of information retrieval (IR) will be prone to using more verbose natural language. Further, through presupposition and direct statements in their interaction with the conversational agent (CA), participants may divulge more personally identifiable information about themselves.

With advancements in natural language inference tools, these PII admissions may be collected to form increasingly accurate user models for use in advertising and accreted by data brokers, which may pose a privacy risk to users of CUI tools.

1 Introduction

1.1 Format

This paper takes the form of a *preregistration report* as described by (Nosek et al., 2018). This is to avoid generating postdictions by describing research goals and methods in advance of performing a study.

1.2 Hypothesis

I hypothesize that study participants performing information retrieval (IR) tasks with a conversational user interface (CUI) will divulge more personally-identifiable information (PII) about themselves than participants interacting with a simple text input field.

Further, I predict that CUIs will encourage participants to use more verbose natural language than if they were presented with a text input field, in turn yielding more opportunities to disclose PII. Responses to the CUI may contain direct disclosures, presuppositions, or no PII.

Moreover, I predict that the verbose natural language encouraged by conversational agents will allow natural language inference (NLI) systems to profile participants more reliably than their peers using text input fields.

1.3 Example

To illustrate our hypothesis, consider the hypothetical scenario that a participant owns a pet bippo[1] and wants to learn what to feed it. I predict that this participant might type a query like *"adult bippo diet"* into a text input field. Due to the conversational nature of many CUIs, I imagine that they might be inclined to phrase their query in more natural language, such as *"What is healthy to feed my adult bippo?"*.

In the former case, there may be a weak signal that the participant owns a bippo, although they may have been a passing interest, or a bippo that they don't own. The latter statement includes presupposition trigger (Karttunen, 1974) through the use of *"my"*, where we understand that the participant has explicitly stated that they own a bippo.

This provides the CUI operator with data for user modeling of participants. This may be accomplished through the use an NLI system to analyze lexical structures present (Kabbara and Cheung, 2022), and attempt to verify the presupposition. (Kim et al., 2021)

1.4 Proposal

I propose a between-group experiment where participants are presented with several information retrieval tasks. The control group is presented with a text input box on a search engine webpage, while the experimental group must conduct their task with a CUI.

The experimental methods are presented in more detail in section 5 and ethical considerations discussed in section 6.

[1] Fictional elements are used in tasks to avoid responses being influenced by preconceptions or sentiment that may otherwise be present that could influence them. See section 3.4 for details.

2 Background

todo: redo this whole section

This research is motivated, in part, by recent announcements from commercial search giants Microsoft (Mehdi, 2023) and Google (Pichai, 2023) to release conversational agents as a new user interface for their search engines. These announcements were presaged by research proposing the integration of generative language models with information retrieval (Metzler et al., 2021), followed by subsequent criticism (Shah and Bender, 2022).

2.1 Business Motivations: User Modeling

Aside from exciting their investors with conversational agents, business has maintained longstanding interest in using natural language processing (NLP) for user modeling (**?**). The financial interest of NLP for user modeling across technology businesses is difficult to understate. By analyzing queries from users, businesses feel they are able to produce more accurate user models, and in turn, send those users advertisements and product recommendations more effectively.

As a case study of this, social media business relies heavily on advertising. By encouraging user engagement and retention, they not only increase the amount of ads seen by users, but perhaps more importantly, they receive a wealth of interactions and content from their users to inform their advertising user models, which are themselves a financial asset (as evidenced not only by advertising platforms, but also data brokers who aggregate user data from across sources).

Central to the hypothesis of this paper is the idea that encouraging users to produce more verbose natural language through CUIs, which provides more robust data for user modeling. If true, this may explain some of the seemingly inexplicable excitement and investment in CUIs recently. While this paper intends to focus on the privacy implications of these interactions, the use of CUIs for user modeling is important to contextualize why CUIs are currently popular[2].

2.2 Business Motivations: Privatization and Regulatory Capture[3]

Reflecting on user modeling by gathering natural language, let's return to the aforementioned Google announcement(Pichai, 2023). Remarkably little is said about why their conversational agent will be used for search, or details about the user-facing CUI itself, but mostly focuses on the large language model (LLM) which will generate the CUI's language.

This may not come as a surprise considering the ongoing interest among established players in the tech industry in creating increasingly large and costly machine learning models. The high operational and development costs associated with the generative LLMs behind recent CUIs might appear counterproductive to an entity whose sole purpose is generating profit, even within the lucrative field of advertising.

The cost of developing LLMs (and other non-language models relying on colossal corpora) has meant research has been driven increasingly large corporations. As profit-motivated entities, one could speculate that driving greater interest in both academia and the industry to use more costly methodology allows these corporations a competitive advantage. The increased barriers to entry make it difficult for competition to develop, but also extends to selling their infrastructure and tools to academics who wish to perform research in these areas.

A recent example of the anti-competitive nature of this can be seen in the testimonies of IBM and OpenAI executives to the US senate (Blumenthal, 2023), urging legislature to enact policy which would require researchers to obtain a license to perform research or develop applications with linguistic models of indeterminate size. Motivations for this testimony may be regulatory capture, which impacts not only their competitors, but also increases cost for researchers in academia who use common statistical models. By increasing barriers to perform research, large corporations may more effectively recruit scholars who require significant sponsorship to obtain a license, use private corpora, and significant compute resources.

[2] I would speculate that the use of CUIs in the tech industry today has few motivations related to usability or functionality, and may be a functional detriment in many places they are applied, but I don't have research to support this claim.

[3] Alternatively titled "Obligatory LLM Section."

2.3 Alternatives Considered

2.3.1 Wizard of Oz

A Wizard of Oz (WOZ) experiment (Jokinen and McTear, 2010) may be useful produce additional insights where the static single-turn tasks proposed in this paper would not be able explore. With an experimenter acting as the conversational agent or search engine, characteristics of multi-turn dialogues can be considered with CAs. This allows for scenarios such as studying how a user may perform a search task when their initial query or question produces unrelated or undesired results.

A WOZ experiment may also be valuable for understanding sociolinguistic aspects of participants interacting with CAs, and allow responses to be tailored to qualitatively explore participant behavior that may not have been initially anticipated in the design of the experiment.

This may raise additional ethical considerations to be aware of. It may be tempting to withhold the information that a human experimenter is playing the role of the conversational agent to produce more accurate-to-life results, this would be a questionable ethical decision. When performing an IRB review, one question that reviewers must ask[4]see: Code of Federal Regulations, Title 45, section 46.111) about a possible deception is whether it would harm or otherwise negatively impact the participant; it is worth considering the impact on participant trust in academic studies once they learn they were deceived.

Additionally, the experiment may leave the participant with impressions about CAs and their use in IR tasks. Whether positive or negative, this may influence their future behavior. In particular, if the experimenter were to produce inaccuracies or disinformation through their participation in IR tasks, it would be paramount to debrief the participant at the end of the experiment and make them aware of any inaccuracies or disinformation they encountered.

2.3.2 Analysis of Real World CA Interactions

A possible followup that I would **not** recommend is one where researcher would be interested in qualitatively analyzing how a participant uses CAs for IR tasks in their day-to-day lives, relatively free of an experimental environment. Whether an experimenter is given access to a search company's corpus of interaction data, or data is collected through a browser extension that monitors CA interactions, this may come with significant privacy and safety concerns for participants.

User data collected en masse may contain significant disclosures of personal information, which if mishandled or compromised, may allow adversaries to de-anonymize a participant. This scenario can cause significant harm to participants, where de-anonymized participants may be harassed, stalked, or otherwise targeted.

It is also unclear if this approach would yield a significant benefit over other experimental methods. The user will still be aware of the study through their consent to monitoring, which may still significantly bias how a user interacts with CAs throughout the study. In this case, it may be equally useful to perform a more controlled study with much lower risk of harm to participants.

3 Methodology

The study takes the form of a between-group experiment where surveys are distributed to each group. A between-group design is used to reduce response bias and variability through demand characteristics in participants. (Rubin and Badea, 2010)

The survey provides information retrieval tasks to participants (see section 3.4, "Tasks") and a search UI [5] to use. The UI given to the control group is text input field on a plain webpage stating it is a search engine, while the experimental group is given a CUI containing a short introduction from the conversational agent (CA) that it is an AI [6] search assistant.

A survey is used for its cost-effectiveness, forgoing the need for experimenters to be present to observe participants, and without needing to perform double-blind testing. Providing participants in each group with identical IR tasks will help ensure that responses are more directly comparable and help eliminate outside variables.

Demographics will be collected through optional questions presented at the end of the survey. See section 3.3 for additional details of demographic data collection.

[4] (

[5] Given sufficient time and resources, the UI may be a view integrated into the survey webpage, where the participant types their query directly into the UI. Otherwise, an image of the UI may be used, followed by the survey text input field.

[6] The term "AI" is used due to its recent popularity as a marketing keyword, and since highly-publicized search CUIs ChatGPT for Bing and Google Bard use this terminology.

3.1 Limitations

A survey is proposed as a time- and cost-efficient method of performing the experiment that forgoes the need for presence of experimenters or double-blind testing. This comes with the limitation of a single-turn discourse for each task, since no experimenter will be present to produce responses relevant to participant inputs. An alternative or future study may wish to pursue a Wizard of Oz design in their experiment to facilitate multi-turn discourse with participants.

An additional limitation is this study does not address speech-based interfaces. Although users will be asked if they used speech-to-text (STT) software to complete the survey, it is not currently a focus of this study, and I expect relatively few participants to use STT. This also neglects the possibility of digital assistants primarily operated by voice commands as an IR interface.

Presuppositions posed by participants in their responses are presumed to be true. This is since tasks use fictional scenarios (see section 4.3), so they cannot be meaningfully verified. In real-world scenarios, CUIs may wish to employ presupposition verification (and to challenge human interlocutors about their false presuppositions) to avoid disseminating disinformation. (Kim et al., 2021)

The survey is will be offered in English, and participants will be asked to write responses in English. This is because I am insufficiently skilled in other languages to perform meaningful analysis on non-English responses, and currently do no have other evaluators to consider responses in other languages.

3.2 Participant Recruitment

As with the rest of this study, participant recruiting efforts must first be approved by IRB. Participants may be sourced from the author's university and workplace via mailing lists. A crowd-sourcing service (e.g. Cint, Prolific, Pavlovia, etc.) may also be considered to recruit additional participants, or participants from specific demographics. This will require funding to pay participants and require the author to review operational details of the platform to ensure ethical treatment of participants and quality of data.

Recruitment materials will describe the study as an investigation of language used when interacting with search interfaces. To avoid revealing the hypothesis to participants (so as to not elicit demand characteristics), but to still responsibly disclose the nature of the study, it should not mention the distinction between experiment groups, or that privacy is the focus of the linguistic analysis.

3.3 Demographic Collection

Optional demographic questions at the end of the survey allow participants to self-identify their age range, gender, ethnicity, computer proficiency, familiarity with CAs[7], and written English proficiency using interfaces described in Table 1.

Free-form text input fields are provided to participants for gender and ethnicity identification in addition to a static selection of common genders to avoid erasing or misrepresenting identities.

Demographic data entered through text input fields is assessed qualitatively, and additional categories for a given demographic selection may be established given sufficient responses identifying with the new category.

3.4 Tasks

Each task is a brief, fictional scenario that prompts participants to perform information retrieval. Both the control and experiment groups are provided the same set of tasks to help ensure responses across can be compared directly, minimizing external variables. Tasks rely on unfamiliar fictional elements to reduce variability from existing preconceptions and sentiments that participants might have.

Likewise, tasks do not assume any background from participants, and does not prompt them for any information outside the fictional scenario. Nevertheless, participants may still share unrelated PII in their responses, which is must be accounted for and handled properly (see section 6).

In attempt to observe how participants may query information for themselves or related to another person, some questions may post fictional scenarios pertaining to the participant directly, while other scenarios involve a friend asking the participant for information or advice. This can impact presupposition triggers (Levinson, 1983) produced where one task may produce more first-person pronouns that form presuppositions (e.g. "**my** dog") and may be useful to observe how these my vary when referring to another (e.g. "my friend's dog", or "their dog") and if participants will have any reticence to mention details of their friend versus themselves to a CA.

[7] "CAs" will be phrased as "AI chatbots" in the survey to accommodate those unburdened by academic jargon.

Demographic	Input type
Age group	Radio selector in 10-year groups
Gender	Selector and text input field (see Appendix A.X)
Ethnicity	Selector and text input field (see Appendix A.Y)
Computer proficiency	Likert scale
Familiarity with CAs	Likert scale
Written English proficiency	Likert scale

Table 1: Optional demographic data questions presented to participants at the end of the survey.

A list of tasks used in this study can be found in appendix A.2. *Note: If this study were to be performed, I will want to find someone with an HCI background to review questions to ensure they are composed in a way that minimizes influence over participant's responses.*

3.4.1 Debrief

At the end of the survey, participants will be asked several questions about the survey and their responses. In particular, participant are asked if there were any questions they were unsure of or confused about, as well as if wish to provide any additional comments or feedback.

Responses here should all be considered, and evaluators should note and categorize any common issues or concerns that participants have. This can then be reported in the study results to help provide context about the data, and to inform researchers of any issues that may impact the data collected.

The end of the survey will include a link to a mailing list they can subscribe to if they wish for updates about the survey. This may help assuage future questions or concerns participants may have, and help engage them in future studies.

4 Results

4.1 Evaluation

Evaluators read each response, noting how many distinct disclosures of PII can be modeled about the user, and whether each disclosure was through a direct statement or a presupposition. Identifying presuppositions is performed based on the presupposition triggers listed by (Levinson, 1983) and (Van Der Sandt, 1992). Additionally, it is noted whether the disclosure is information pertaining to the participant, or to someone else.

For example, a response such as *"What should I feed* ***my*** *pet bippo?"* presupposes that the participant has a pet bippo and is be classified as a disclosure. Likewise, *"I have a bippo. What do I feed it?"* contains a direct statement and is a disclosure. By contrast, *"What do bippos eat?"* a query *"bippo food"* implies that the participant is interested in the diet of a bippo, and may weakly suggest ownership of a bippo, but with little certainty. As such, it is not tallied as a disclosure.

I am currently the only and evaluator, but if interested parties assist with evaluation of study data, then evaluations can be cross-validated.

4.2 Data Validation

Although text fields in the survey are intended for tasks, a participant may misunderstand and describe what actions they would take, rather than typing their literal search. Unless this description provides a clear quote of what they would enter into the interface, this response would need to be discarded.

Likewise, it is possible participant may also express their confusion to a task in the text response, or otherwise provide remarks or commentary on the task (e.g. *"I don't know what I'm supposed to do here"*, or *"I would not perform a search for this"*

4.2.1 Crowd-Sourcing Platforms

If a crowd-sourcing platform is used, data my need to be validated in advance to identify invalid surveys. For example, if an identical value is pasted in every text field, or if the entire survey is completed in a few seconds, it may indicate a disinterested participant who has not read the questions. Qualitative analysis may also be employed; if a response appears to be entirely unrelated to the task or completely nonsensical, it may indicate that the response was generated.

Invalid responses will not be used in analysis. Depending on the features provided by the crowd-sourcing platform, participant who produced invalid results may be excluded from future studies. It should be noted that even if a participant generated invalid responses, that **they must still be paid for their work** as an ethical consideration.

4.3 Analysis

Note: Unfortunately the UW statistics consulting was booked for the Spring term, so I was unable to undergo a stats review in time for this paper.

4.3.1 Hypothesis Testing

Since some participants may be more or less prone to disclose PII in their responses, find the rate of disclosure for each participant considering all their responses. This rate is calculated for each category of disclosure: directly stated disclosures, disclosure through presupposition, and no disclosure.

Then a contingency table is constructed using the three categories of disclosure rates as one axis and the control and experiment groups as the other. Given the contingency table, a chi-squared test is applied to the distribution for hypothesis testing.

Hypothesis Test Contingency Table		
Disclosure	Control	Experiment
Direct		
Presupp.		
None		

4.3.2 Demographic Analysis

Demographic data is analyzed similarly, by forming a chi-squared distribution for each demographic. Compared to the hypothesis test, there are now tuples of the control and experiment group columns, pairing them each with a demographic subcategory[8].

Confidence intervals are calculated and visualized for the distributions and compared. for demographic categories to understand the possible variability from small demographic samples.

In cases where there are fewer than 3 respondents in a particular demographic, the sample will be deemed too small, and analysis will be skipped for this group.

4.4 Qualitative Analysis

Responses should each be read and considered by evaluators, who should note and categorize any trends that may lend insight into interactions with CAs for future work, as well as provide context for the results reported by this study.

Debrief questions (see: section 3.4.1) should also be analyzed, noting common issues, concerns, and sentiments from users. As above, this may provide insight or raise additional questions for future work. If participants commonly express confusion or uncertainty about the tasks they performed, it may indicate that the data may not reflect real interactions or that data may contain statistical bias as a result. This should be accounted for when considering the validity of results, as well as to help guide future research away from recreating the same mistakes encountered in this study.

5 Discussion

It the hypothesis is true, industry actors using CUIs may have the opportunity to more accurately model users for advertising, personalization, and marketing research. This comes at the cost of privacy for people who use the CUIs, who may be subtly influenced to use language that divulges more information about themselves. This could be exacerbated if a CUI is the only interface available for a given application.

Recent research suggests that this use of CAs may not be appropriate for IR tasks (Shah and Bender, 2022). As such, people needing to use a CUI IR application may experience a detriment in both search quality, and in personal privacy.

Additionally, while corporations using CUIs for IRs may view the additional PII as a benefit to their advertising models, these benefits must be weighed realistically against both the cost to develop and run them, plus the cost of handling PII correctly across jurisdictions.

If this hypothesis is true, it would appear that CAs may be even less appropriate for general IR tasks than research has already suggested and may come at the mutual detriment of those who deploy them and those who must use them.

6 Ethical Considerations

Since this study concerns PII, and since users may disclose actual PII into the text input forms provided to them, this study must ensure the privacy of its participants.

During recruitment, any identifiers (whether through a crowd-sourcing service, or through mailing list memberships) must only be retained long enough to ensure that participants aren't distributed repeated surveys. Unique identifiers in this process should be replaced with a UUID. A hash of a participant's name and email may be sufficient for a UUID, since participant information cannot be

[8] e.g. for the Age demographic, it may have columns [control group, 21-30yrs], [experiment group, 21-30yrs], [control group, 31-40], and so on.

deduced from the hash, but the hash may be used to check if a potential participant has already been sent a survey.

Participant names and emails are not solicited in the survey itself. To reduce the likelihood of correlating any response to any particular identity, demographic responses and tasks responses will be stored and analyzed separately, and any identifying fields (including UUIDs) should not be revealed to evaluators when performing analysis. (Jokinen and McTear, 2010)

If a participant's response contains PII about themselves or others (distinct from the fictional task scenarios), the PII response should be anonymized before publishing any data or reports in order to protect the privacy of participants. An index of which results have been anonymized maintained, noting which words are substitutes inserted by the evaluator. In the event a result cannot be meaningfully anonymized, it should be excluded from the analysis and reports, making note of how many responses were excluded in this way.

To emphasize a point from section 4.2 on data validation, when a crowd-sourcing platform is used, we must ensure that all participants are paid for their work, even if they produce invalid responses.

7 Conclusion

Motivated by recent popularity of generative language CUIs, this paper proposes a between-group study of language used by people interacting with CUIs. By analyzing both direct statements and presuppositions from respondents, I hope to answer the hypothesis that people using CUIs for IR may be more inclined to divulge PII compared to search engines using a simple text input field for queries.

References

Richard Blumenthal. 2023. Oversight of A.I.: Rules for Artificial Intelligence.

Kristiina Jokinen and Michael McTear. 2010. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lauri Karttunen. 1974. Presupposition and Linguistic Context. 1(1-3).

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945. Association for Computational Linguistics.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Yusuf Mehdi. 2023. Reinventing search with a new ai-powered microsoft bing and edge, your copilot for the web.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: Making domain experts out of dilettantes. 55(1):1–27.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. 115(11):2600–2606.

Sundar Pichai. 2023. An important next step on our AI journey.

Mark Rubin and Constantina Badea. 2010. The central tendency of a social group can affect ratings of its intragroup variability in the absence of social identity concerns. 46(2):410–415.

Chirag Shah and Emily M. Bender. 2022. Situating Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 221–232. ACM.

Rob A. Van Der Sandt. 1992. Presupposition Projection as Anaphora Resolution. 9(4):333–377.

A Appendix

A.1 Survey introduction

Introductory text to the survey is listed below.

"This survey will present several information retrieval tasks and a picture of a web search tool.

Type into the answer box what you would write into the pictured web search tool to solve the task. If you would not use this tool to perform the given task, please note it in the answer box. Data from this survey will be used to develop an upcoming study. There is no right or wrong answer. If possible, please write responses in English."

See also appendix A.4 for a mockup of the UI including the above text.

A.2 Tasks

As noted in section 3.4, if this study is performed, someone with HCI expertise should first be consulted to mitigate the risk of biasing user responses through phrasing of questions. Toward this, the tasks are described generally, and participants

TODO: Write more tasks

1. "You have a new pet bippo, but you don't know what to feed it. What would you type into the tool picture below?"
2. "A new president has been elected in the country of Bippopolis, but you don't know the president's name. What would you type into the tool picture below?"
3. "A Wug has nested in your friend's house. Unhappy about this, they ask you for advice. What would you type into the tool picture below?"
4. "You heard that legislation is being debated to forbid Wugs in homes. What would you type into the tool picture below?"

A.3 UI mock-ups

Note: Mockups need to be updated slightly to reflect revised text described in the paper.

Note: CUI mockup needs to look like like ChatGPT.

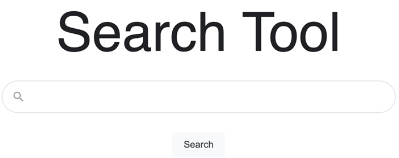


Figure 1: UI mockup for control group

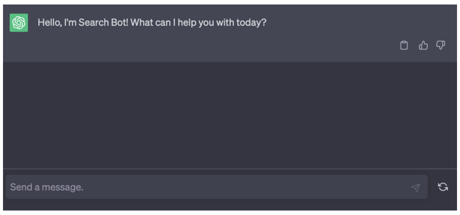
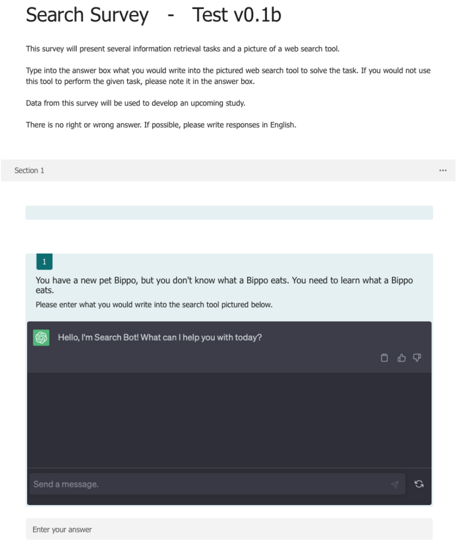

Hello, I'm Search Bot! What can I help you with today?

Figure 2: UI mockup for experiment group

A.4 Survey UI mockup


Search Survey - Test v0.1b

This survey will present several information retrieval tasks and a picture of a web search tool.

Type into the answer box what you would write into the pictured web search tool to solve the task. If you would not use this tool to perform the given task, please note it in the answer box.

Data from this survey will be used to develop an upcoming study.

There is no right or wrong answer. If possible, please write responses in English.

Section 1

1

You have a new pet Bippo, but you don't know what a Bippo eats. You need to learn what a Bippo eats.

Please enter what you would write into the search tool pictured below.

Hello, I'm Search Bot! What can I help you with today?

Figure 3: UI mockup for the survey using Microsoft forms.