Evaluation Sheet for Deep Research: A Use Case for Academic Survey Writing

Israel Abebe Azime[1,*] **, Tadesse Destaw Belay**[2,*]**, Atnafu Lambebo Tonja**[3,*]
[1] Saarland University, [2] Instituto Politécnico Nacional, [3] MBZUAI

Abstract

Large Language Models (LLMs) powered with argentic capabilities are able to do knowledge-intensive tasks without human involvement. A prime example of this tool is Deep research with the capability to browse the web, extract information and generate multi-page reports. In this work, we introduce an evaluation sheet that can be used for assessing the capability of Deep Research tools. In addition, we selected academic survey writing as a use case task and evaluated output reports based on the evaluation sheet we introduced. Our findings show the need to have carefully crafted evaluation standards. The evaluation done on OpenAI's Deep Search and Google's Deep Search in generating an academic survey showed the huge gap between search engines and standalone Deep Research tools, as well as the shortcomings in representing the targeted area.

1 Introduction

Deep Research tools are designed to create comprehensive, long-form reports that dive deep into complex topics (Wu et al., 2025). Their defining characteristics include unassisted web browsing, compilation of several sources, long waiting time, and results that resemble reports, not chat responses (OpenAI, 2025). Deep Research improves traditional search capabilities from keyword-based searching to more exhaustive search incorporating reasoning, inference synthesis, and response generation. This profound research feature transcends basic question-answering; it enables LLMs to navigate the internet, process extensive datasets, synthesize insights, and create structured reports with appropriate citations (Xiong et al., 2024).

LLM providers such as Google[1], OpenAI[2], Perplexity[3], XAI[4], and others are making available their Deep Research agent-based applications.

Deep Research tools are increasingly used to assist academic tasks like literature reviews, offering draft summaries in minutes and aggregating data from numerous sources. However, they still require oversight, as they may hallucinate, cite unreliable sources, or prioritize outdated content. Even though, Deep Research tools are powerful for scaling up our research capabilities, users must understand their strengths and limitations to choose the right tool. In this work: **1)** We introduce ***Evaluation Sheet*** as a road-map for evaluating the performance of Deep Research tools. **2)**As a use case (intended only as an example), we selected three recent NLP survey papers focused on African countries and languages: an Ethiopian language survey (Tonja et al., 2023), a Nigerian language survey (Inuwa-Dutse, 2025), and a Kenyan language survey (Amol et al., 2024) to assess the applicability of the introduced evaluation sheet in order to evaluate the generated Deep Research report.

2 The Evaluation Sheets - Pillars

LLM evaluation datasets, particularly those focusing on low-resource languages, should emphasize specific characteristics of the generated output. In this work, we propose evaluation sheets that contain different questions in five pillars to evaluate LLMs' Deep Research tool

(1) LLMs & Deep Research for [*Surveying NLP Papers and Datasets for Low-Resource African Languages*][5] Surveying existing NLP papers in research areas such as low-resource languages presents unique challenges. A crucial task

[*] Equal Contribution.

[1] https://blog.google/products/gemini/google-gemini-deep-research/

[2] https://openai.com/index/introducing-deep-research/

[3] https://www.perplexity.ai/ko/hub/blog/introducing-perplexity-deep-research

[4] https://x.ai/blog/grok-3

[5] This section and subsequent questions can be replaced or modified according to the use case scenario (e.g., gender bias analysis, linguistic inclusion, or indigenous language documentation).

is determining whether these tools can effectively identify the most important and impactful research, even when such research papers do not appear in the top search results. The primary issue we aim to address is how the growing popularity of these tools and their increasing role in replacing traditional searche engines affects the ***visibility and accessibility of significant research***.

(2) Hallucination Hallucination refers to information that appears true to someone without prior knowledge of the subject but cannot be verified by a reliable source (Huang et al., 2025). In contrast, errors are categorized as mistakes that are easily noticeable. Hallucination is a huge treat in practical LLM usage, specifically while automating knowledge extraction from contents like research works. This set of guidelines and questions helps us determine the focus we must place on the reliability of the output.

(3) Correctness of sources Sources can range from reliable, peer-reviewed papers to blogs and social media pages that present personal opinions. While extracting information from both types of sources is optional, web agents should be able to distinguish between reliable and unreliable sources.

(4) Information Validity The validity of the references provided can be assessed based on their accessibility, verification through independent sources, and whether they demonstrate why they are superior to other potential alternatives.

(5) Information Latest-ness Recent information is more valid compared to older information that may have a high search volume but could have been corrected or improved by more recent works. Research papers with higher citation counts and those that appear at the top of search results are not always the latest studies, which can pose a challenge for LLM agents searching the web for information.

(6) Quantifying Actual Google Search Results vs. Deep Research Answers

Finally, we added questions to explore how the shift from using search engines like Google for information retrieval compares to using automated search agents like Deep Research tools.

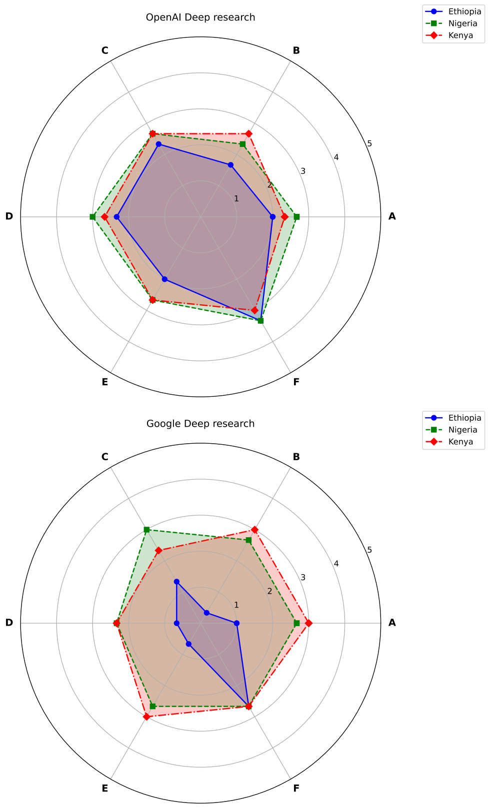

Figure 1: **A** – LLMs & Deep Research for Surveying NLP Papers, **B** – Hallucination, **C** – Correction Sources, **D** – Information/Link Validity, **E** – Information Latestness, **F** – Quantifying Actual Google Search Results vs. LLM Answers,

3 Case study: Ethiopia, Nigeria, Kenya

3.1 Methodology

Creating evaluation sheet We selected three regional survey papers that focus on capturing valuable research progress within their respective countries: the Ethiopian language survey (Tonja et al., 2023), the Nigerian language survey (Inuwa-Dutse, 2025), and the Kenyan language survey (Amol et al., 2024). We analyzed these papers in detail, extracted the key questions they addressed, and then combined them to formulate prompts (see D)incorporating these questions. To create the evaluation sheet, we carefully identified scenarios the Deep Research tools fail at and must be tested with and created a list of questions under each important evaluation topic.

Generating representative outputs We evaluated the prompts for validity and selected the one capable of generating detailed reports. Using a selected prompt, we generated three distinct Deep Research outputs by modifying only the country-specific information while utilizing OpenAI Deep Research and Google Deep Research. Three reviewers selected from the authors of this study reviewed the outputs of the tools and rated the generated report based on the rating criteria for each question in the pillars. They used the actual research paper from each of the countries as a reference while answering the questions accordingly.

3.2 Comparative analysis

In this section, we discuss our observations while evaluating reports generated by Google's Deep Research and OpenAI's Deep Research tools.

LLMs & Deep Research for Surveying NLP Papers Both Google's Deep Research and OpenAI's Deep Research tools show below-average results in identifying more valuable research works in their reports. The region-specific gap becomes larger for Google's Deep Research.

Hallucination The inclusion of social media links alongside verified academic peer review catalogs as sources makes Deep Research tools particularly susceptible to hallucinations and erroneous outputs. Additionally, the absence of source information in reports or the citation of incorrect sources complicates the process of identifying and verifying hallucinations. However, based on our analysis, we found that the rate of misinformation and hallucination is not significantly high.

Correctness of Sources When examining the detailed process these tools follow while "researching", they tend to review a large number of relevant resources. Google's tool heavily summarizes information and often does not mention many of the sources it picks up during the process. Additionally, both tools tend to include social media links, such as Facebook and Reddit, as sources of information.

Information/Link Validity We observe that the tools use sources multiple times during their execution. Apart from that, the tools have a problem of identifying the correct source from which the information is obtained and mostly rely on survey papers and summarized contents rather than extracting information from the original source.

Actual Google Search Results vs. LLM Answers Although the system does not produce significant misinformation, its outputs are not fully aligned with Google search results. We find better choices, more recent works, and broader domain coverage when using Google Search.

3.3 Lesson learned - Takeaway

The need for evaluation standard With the rapid introduction of tools that improve or entirely replace search engines, it is crucial to establish evaluation guidelines that foster consistency and common characteristics across benchmarks. The careful design and assessment of these tools are essential, as they shape the knowledge and research considered important, as well as how different approaches and solutions are presented for comparison, ultimately influencing decision-making. If these tools are not designed to provide as much relevant information as possible to users, the real decision-making process, including the selection of problems and solutions, risks being controlled by autonomous agents developed by big tech companies.

Are Deep Research tools reliable for extracting information and generating user-ready reports for low-resource research summarization? The use cases in this study, focused on generating scientific summary reports on underrepresented groups, highlight the challenges of finding, sorting, and presenting hard-to-access research. We found that Deep Research tools are **not fully reliable**, as their selection of research works lacks transparency, and their summaries, drawn from multiple sources,fail to comprehensively represent the research landscape of the targeted area.

Despite the limitations discussed above, Deep Research tools have the potential to present summarized information and make it more accessible.

4 Conclusion

In this work, we developed an Evaluation Sheet to help researchers identify the most critical evaluation criteria for assessing Deep Research tools for different use cases. This evaluation sheet seeks to standardize benchmarking datasets by highlighting key focus areas. To demonstrate its applicability, we conducted a proof-of-concept study on "Deep Research for Survey Paper Generation" and used it to evaluate two well-known Deep Research tools.

We hope researchers will adopt this Evaluation Sheet to create benchmarking datasets in their respective domains, ultimately improving the effectiveness of agentic tools that require minimal human interaction. By ensuring these tools generate reliable and informative outputs comparable to those found through independent searches, we aim to enhance their practical utility and trustworthiness.

Limitation

Deep Research tools are relatively new, and we selected OpenAI and Google as use cases due to their availability and popularity. Future research will expand the scope by incorporating a broader range of tools, generating a larger number of reports and a larger number of evaluators to better assess their capabilities on a wider scale.

Acknowledgment

The authors would like to thank the German Federal Ministry of Education and Research and the German federal states (http://www.nhr-verein.de/en/our-partners) for supporting this work/project as part of the National High-Performance Computing (NHR) joint funding program.

References

Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Caringtone Kinyanjui, Steven Ndung'u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. State of nlp in kenya: A survey.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Isa Inuwa-Dutse. 2025. Naijanlp: A survey of nigerian low-resource languages.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

OpenAI. 2025. Deep research system card.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in Ethiopian languages: Current state, challenges, and opportunities. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 126–139, Dubrovnik, Croatia. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges.

Appendix

A What are deep reserch tools?

Unlike traditional search engines, which primarily provide direct answers, it employs an iterative search process that deconstructs complex inquiries and engages in reasoning before generating responses (Wu et al., 2025). This method operates several search cycles, such as an iterative reading, searching, and reasoning cycle, until the most accurate response is achieved. The entire operation can be segmented into three main distinct phases (search, read and reason), as illustrated in Figure 2.

B The Evaluation Sheets - Pillars

(1) LLMs & Deep Research for [*Surveying NLP Papers and Datasets for Low-Resource African Languages*][6]. Surveying existing NLP papers in research areas such as low-resource languages presents unique challenges. A crucial task is determining whether these tools can effectively identify the most important and impactful research, even

[6]This section and subsequent questions can be replaced or modified according to the use case scenario (Eg. financial market study, Sport analysis etc).

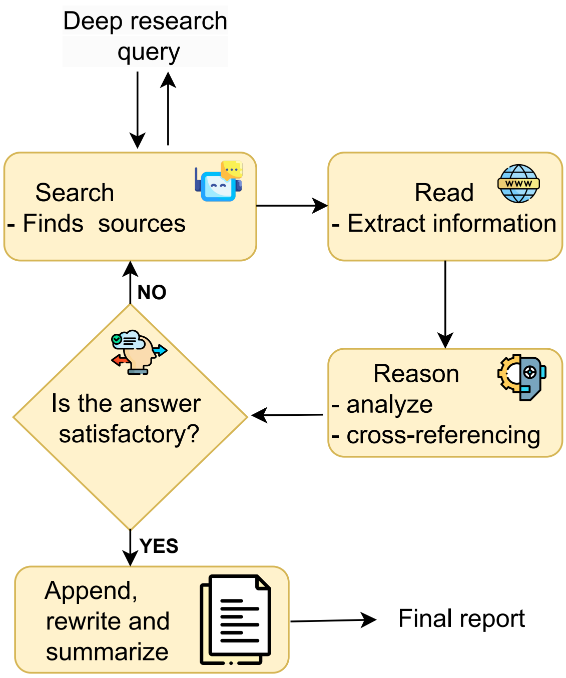

Figure 2: Deep Research workflow

when such research papers do not appear in the top search results. The primary issue we aim to address is how the growing popularity of these tools and their increasing role in replacing traditional searche engines affects the ***visibility and accessibility of significant research***.

To access the usage of LLMs & Deep Research in survey report writing in low-resource languages, we crafted the following question:

- *Does the Deep Research reports effectively identifies and consolidate NLP papers on low-resource [African][7] languages?*
- *Does the selection of datasets for low-resource [African] languages is comprehensive and representative?*
- *Does the Deep Research method provide sufficient depth in its analysis of linguistic challenges in [African] NLP?*
- *Does the LLM-generated survey highlights the most impactful research in [African] NLP?*
- *Does the coverage of low-resource [African] languages in the survey align with the actual research landscape?*

(2) Hallucination Hallucination refers to information that appears true to someone without prior knowledge of the subject but cannot be verified by a reliable source (Huang et al., 2025). In contrast, errors are categorized as mistakes that are easily noticeable. Hallucination is a huge treat in practical LLM usage, specifically while automating knowledge extraction from contents like research works. This set of guidelines and questions helps us determine the focus we must place on the reliability of the output. The following questions are crafted to evaluate whether the Deep Research generated report contains hallucination.

- *Does the Deep Research generated survey contains minimal factual errors or hallucinations?*
- *Does the hallucinated content, if present, is easy to identify and correct?*
- *Does the Deep Research tool properly distinguishes between verified academic sources and speculative content?*
- *Does a lower risk of hallucination improve the reliability of the survey's insights?*

(3) Correctness of sources Sources can range from reliable, peer-reviewed papers to blogs and social media pages that present personal opinions. While extracting information from both types of sources is optional, web agents should be able to distinguish between reliable and unreliable sources. Below, we pose a set of questions to assess whether the source impacts the reliability of the information and whether certain sources are preferable. This approach ensures that the extracted information is accurate and verified.

- *Does the sources suggested in the report are based on verifiable and authoritative sources?*
- *Does the Deep Research tool appropriately prioritize papers on credibility and impact?*
- *Does the mechanism used by Deep Research to extract information from sources adequately account for domain-specific knowledge in [NLP]?*

(4) Information Validity The validity of the references provided can be assessed based on their accessibility, verification through independent sources, and whether they demonstrate why they are superior to other potential alternatives. Below are the questions created to assess the validity of information generated by Deep Research.

- *Does the cited links and references in the survey are valid and accessible?*
- *Does the Deep Research tool effectively differentiates between credible and non-credible sources?*
- *Does the report content remains valid and relevant when cross-checked with independent sources?*
- *Does the Deep Research tool provide sufficient transparency regarding how sources are selected and ranked?*
- *Does the Deep Research generated report appropriately handles broken or outdated links in its output?*

[7] can be specific region name (Ethiopia, Kenya and Nigeria)

(5) Information Latestness Recent information is more valid compared to older information that may have a high search volume but could have been corrected or improved by more recent works. Research papers with higher citation counts and those that appear at the top of search results are not always the latest studies, which can pose a challenge for LLM agents searching the web for information. The following question will help to assess whether the information generated in the report has been extracted from the latest sources.

- *Does the report prioritize the most recent sources?*
- *Does the Deep Research tool effectively identify the latest trends in NLP for low-resource African languages?*
- *Does the Deep Research method ensure that outdated references are minimized in the survey?*
- *Does the system effectively highlight emerging resources that are not widely recognized?*
- *Does the report output remain relevant given the fast-paced evolution of AI and [NLP] research?*

(6) Quantifying Actual Google Search Results vs. Deep Research Answers

Finally, we added questions below to explore how the shift from using search engines like Google for information retrieval compares to using automated search agents like Deep Research tools.

- *Does the report findings align well with actual Google search results on the same topics?*
- *Does Deep Research generated answers provided by Deep Research are insightful than Google search results?*
- *Does the Deep Research tool accurately quantify differences in retrieval efficiency between LLMs and traditional search engines?*
- *Does the Deep Research tool effectively reduce misinformation compared to open-web search engines?*
- *Does the Deep Research approach provide added value beyond standard keyword-based search queries?*

C Rating Procedure

For the above questions (listed in Section 2), we recommend that users use the Likert scale (Joshi et al., 2015) rating system when answering. The rating scale consists of six levels to express agreement or disagreement with a question. These are: **Strongly Disagree (0)**- indicates complete opposition with no support for the statement. **Disagree (1)**- reflects mostly disagreement, though some merit is acknowledged. **Somewhat Disagree (2)**- suggests a leaning toward disagreement while recognizing certain validity. **Neutral (3)**- signifies neither agreement nor disagreement or an undecided stance. **Somewhat Agree (4)**- represents general agreement but with some reservations. Finally, **Strongly Agree (5)**-expresses full endorsement and support without any doubt.

D Prompt

Deep Research Template for NLP Survey on a Specific Country

Steps to Conduct This NLP Survey

Step 1: Define Your Research Scope Select the country whose NLP landscape you want to analyze. Identify the languages spoken in the country, including official, regional, indigenous, and endangered languages. Decide on the specific NLP focus, such as general NLP, speech recognition, machine translation, or sentiment analysis.

Step 2: Gather Data & Sources

- **Academic Papers**: Search IEEE Xplore, ACL Anthology, Google Scholar, arXiv, and Scopus.
- **Datasets & Resources**: Explore Hugging Face, Kaggle, LDC, and government data repositories.
- **Pretrained Models**: Check models from Hugging Face, Google AI, and Meta AI.
- **Government & Industry Reports**: Look for language policy documents and AI research reports.
- **Community & Open-Source Projects**: Identify ongoing grassroots NLP efforts.

Step 3: Structure the Paper Using the Template Below

Use the structured sections to analyze and organize findings. Answer the guiding questions within each section to provide a comprehensive analysis.

Step 4: Conduct Systematic Analysis

Review historical NLP progress in the country. Evaluate language challenges and computational constraints affecting NLP adoption. Identify key gaps in linguistic resources, datasets, and models. Highlight ongoing projects and promising research directions.

Step 5: Synthesize Findings & Propose Solutions

Summarize research trends, NLP applications, and linguistic barriers. Suggest data collection initiatives, model improvements, and collaborative strategies. Provide policy recommendations for governments, industries, and researchers.

Research Template: Structure of the Paper

- **Introduction**: Define the research focus, its importance, and the major linguistic and computational challenges in the country.
- **Research Methodology**: Describe the sources used, search strategies, and inclusion/exclusion criteria.
- **Language Landscape**: Analyze linguistic diversity, digital presence, and computational challenges.
- **Available NLP Resources & Tools**: Review datasets, pretrained models, and language processing tools.
- **NLP Applications & Downstream Tasks**: Discuss various NLP tasks such as text processing, machine translation, ASR, NER, and conversational AI.
- **Challenges & Limitations**: Address technical constraints, linguistic barriers, and ethical concerns.
- **Future Directions & Recommendations**: Propose solutions for data collection, model improvements, policy considerations, and community engagement.
- **Conclusion**: Summarize key findings and provide a call to action.

Guiding Questions for Each Section 1. Introduction

- What is the focus of this research?

- Why is this topic important for [Country Name]?
- What are the major linguistic and computational challenges in this country's NLP landscape?
- What are the objectives and scope of this study?
- How does the country's NLP research compare to global trends?

2. **Research Methodology**
   - What databases and sources were used?
   - What search strategies were applied?
   - What criteria were used to include/exclude studies?
   - How was the information categorized (e.g., by language type, NLP task, dataset availability)?

3. **Language Landscape in [Country Name]**
   - What are the primary linguistic characteristics of the country's languages?
   - Which languages have the most NLP research, and which are neglected?
   - What challenges arise in processing these languages (e.g., word segmentation, diacritics)?

4. **Available NLP Resources & Tools**
   - Are there high-quality datasets available for these languages?
   - Are the models pre-trained on country-specific linguistic data?
   - What tools exist for POS tagging, NER, and other NLP tasks?

5. **NLP Applications & Downstream Tasks**
   - What NLP tasks have seen the most research focus?
   - What tools and datasets exist for these tasks?
   - What are the biggest challenges in implementing NLP solutions?

6. **Challenges & Limitations**
   - What are the biggest challenges preventing NLP advancements?
   - Are there systematic biases in datasets and models?
   - How does governmental or industry support impact NLP growth?

7. **Future Directions & Recommendations**
   - What strategies can bridge the research gap in NLP for [Country Name]?
   - What government or private sector initiatives can support NLP growth?
   - How can the NLP community collaborate to improve datasets and models?

8. Conclusion
Summarize key findings and provide a call to action for researchers, policymakers, and industry leaders.
Practical Example: Applying This Template

- **Choose the country**: Kenya.

- **Select the languages**: Swahili (major language), Kikuyu, Luo, Maasai (regional languages).
- **Determine the focus**: Speech recognition & machine translation.
- **Collect data**: Look for Kenyan NLP research, datasets, and community projects.
- **Analyze findings**: Identify gaps, challenges, and progress in NLP research.
- **Suggest solutions**: Recommend better dataset collection, funding initiatives, and collaborative research.

E Results

Category	**Criteria**	**openai**			**google**		
		Ethiopia	Nigeria	kenya	Ethiopia	Nigeria	kenya
LLMs & Deep Research for Surveying NLP Papers and Datasets for Low-Resource African Languages	The surveyed LLMs effectively identify and consolidate NLP papers on low-resource African languages.	2.00	2.67	2.33	1.00	2.67	3.00
	The selection of datasets for low-resource African languages is comprehensive and representative.	1.67	2.33	2.67	0.33	2.67	3.00
	The deep research method provides sufficient depth in its analysis of linguistic challenges in African NLP.	2.33	2.67	2.67	1.33	3.00	2.33
	The LLM-generated survey highlights the most impactful research in African NLP.	2.33	3.00	2.67	0.67	2.33	2.33
	The coverage of low-resource African languages in the survey aligns with the actual research landscape.	2.00	2.67	2.67	0.67	2.67	3.00
Hallucination	The LLM-generated survey contains minimal factual errors or hallucinations	3.33	3.33	3.00	2.67	2.67	2.67
	The hallucinated content, if present, is easy to identify and correct.	2.33	2.33	2.33	2.33	2.33	2.67
	The AI system properly distinguishes between verified academic sources and speculative content.	2.67	2.67	2.67	2.00	2.00	1.67
	The risk of hallucination significantly impacts the reliability of the survey's insights.	2.67	2.33	2.33	1.33	1.67	1.67
Correction Sources	The papers suggested in the survey are based on verifiable and authoritative sources.	1.67	2.33	2.33	2.33	2.33	2.33
	The correction process effectively improves the reliability of the final survey report.	2.00	2.00	2.33	1.67	2.00	2.00
	The AI system appropriately prioritizes papers on credibility and impact.	2.00	1.67	2.00	1.00	2.33	2.00
	The mechanism used by deep research to extract information from papers adequately account for domain-specific knowledge in NLP.	2.33	2.33	2.67	1.67	2.67	2.33
Information/Link Validity	The cited links and references in the survey are valid and accessible.	2.00	1.50	2.00	3.00	3.67	3.50
	The AI effectively differentiates between credible and non-credible sources.	1.67	2.00	2.33	1.67	2.33	2.33
	The survey content remains valid and relevant when cross-checked with independent sources.	2.00	2.67	2.33	2.33	2.00	3.00
	The system provides sufficient transparency regarding how sources are selected and ranked.	1.00	1.00	1.00	1.33	1.33	1.33
	The AI-generated survey appropriately handles broken or outdated links in its output.	1.33	1.67	1.67	1.67	1.33	1.67
Information Latestness	The survey prioritizes the most recent research papers and datasets.	2.67	2.67	2.67	1.33	2.33	2.00
	The AI system effectively identifies the latest trends in NLP for low-resource African languages.	2.67	2.67	3.00	1.67	2.33	2.67
	The deep research method ensures that outdated references are minimized in the survey.	1.67	2.33	2.33	1.67	2.33	2.00
	The system effectively highlights emerging datasets that are not widely recognized.	2.00	1.67	1.33	0.33	2.00	1.67
	The survey output remains relevant given the fast-paced evolution of AI and NLP research.	2.67	2.67	2.67	1.33	2.00	2.00
Quantifying Actual Google Search Results vs. LLM Answers	The survey findings align well with actual Google search results on the same topics.	2.00	2.67	3.00	1.33	2.67	2.67
	LLM-generated answers provided by deep research are insightful than Google search results.	2.33	2.67	2.67	1.00	1.67	1.67
	The AI system accurately quantifies differences in retrieval efficiency between LLMs and traditional search engines.	1.50	2.00	1.50	2.00	2.00	2.00
	The system effectively reduces misinformation compared to open-web search engines.	3.00	2.67	3.00	2.00	1.67	1.67
	The deep research approach provides added value beyond standard keyword-based search queries.	2.67	2.33	2.67	2.00	1.33	2.00

Table 1: Labeling results shown in 1