Human Interaction and Emerging Technologies (IHIET-AI 2024), Vol. 120, 2024, 172–185
https://doi.org/10.54941/ahfe1004567

TauchiGPT_V2: An Offline Agent-Based Opensource AI Tool Designed to Assist in Academic Research

Ahmed Farooq[1], Jari Kangas[1], Mounia Ziat[2], and Roope Raisamo[1]

[1]Tampere Unit of Computer Human Interaction, Tampere University, Tampere, Finland
[2]Information Design and Corporate Communication department, Bentley University, Waltham, MA 02452, USA

ABSTRACT

Recent progress in artificial intelligence, particularly deep learning, has ushered in a new era of autonomously generated content spanning text, audio, and visuals. This means Large Language Models (LLMs) such as ChatGPT, Llama2, Claude, and PaLM 2 are now developed enough to not only fill in the gaps within user-generated content, but also create unique content of their own, using predefined styles, formats, and writing techniques. With selective modelling and fine-tuning relevant training data, LLMs can output original content for a wide range of tasks previously considered solely the domain of human creativity. However, if we look at the area of research and development within academics, this AI renaissance has yet to make a meaningful impact finding in the pedagogical domains. Crafting a tailored R&D instrument, adept at intricate research procedures, previously presented a formidable challenge regarding expertise, time, and fiscal resources. However, the latest development within this context, Generative Pre-trained Transformers (GPT) and their foundational structures offer a beacon, given their potential to exploit pre-trained Large Language Models (LLMs) for optimizing standard research operations. Our previous work on Autonomous Agents shows that using existing tools and deductive reasoning techniques built on the LangChain model can create a customized tool for academic research. This study builds on the existing work in autonomous agents and open-source LLMs to develop TAUCHI-GPT_V2, a novel adaptation of the academic research assistant. TAUCHI-GPT_V2, conceptualized as an open-source initiative, is built on top of the LangChain architecture employing LLaMA2-13b as the core LLM, ingesting users' own data and files to provide highly relevant contextual results. In this paper, we discuss how TAUCHI-GPT_V2 uses custom offline localized vectorDB for parsing users' personal files to output relevant contextual results within a chat interface. We also put the model to the test by having academic researchers utilize the tool within their daily workflow and report its efficacy and reliability in both hallucinations as well as citing relevant information to enhance user workflow for academic research-related tasks.

Keywords: Artificial intelligence, Generative pre-trained transformers, Autonomous agents, Large language models (LLMS), ChatGPT, Chain of though (COT), Tree of thought (TOT), AI alignment, Human computer interaction, Research and development, Foundation AI models, Responsible AI (RAI)

INTRODUCTION

Over the past few years, most disciplines have experienced significant shifts, primarily due to technological integrations in Artificial intelligence (AI) and machine learning. Rapidly advancing capabilities in natural language processing (NLP) have stood out as a pivotal force reshaping most disciplines. AI's proficiency in comprehending, decoding, and producing human language has unveiled novel opportunities for enriching academic research and how information is shared within research groups. This article delves into the growing domain of AI and NLP, exploring their roles in scholarly pursuits. As the academic world continually searches for innovative tools to boost the spread and absorption of knowledge, AI has paved the way for an array of resources tailored for literature reviews, content scrutiny, and scientific authoring. These advancements not only optimize academic workflows but also enrich the depth of scholarly content.

Nonetheless, incorporating online AI tools into scholarly endeavours carries inherent ethical dilemmas. Issues surrounding authenticity, data privacy, and potential biases are paramount. Therefore, it is also essential to develop and research tools that work locally on users' machines, incorporating their own outputs to customize and generate relevant and consistently reliable outputs necessary for various academic workflows. Furthermore, researchers must strike a balance between leveraging AI utilities and preserving their unique knowledge, writing flair, and inventive spark, ensuring AI serves as an aid rather than a replacement.

RELATED WORK

The paradigm of Human-Centred Artificial Intelligence (HCAI) emphasizes prioritizing human needs over mere technological advancements in AI (Bingley et al., 2023). As AI continues to make strides, there's an increasing impetus to ensure that it augments rather than supplants human abilities. Distinct projects such as the European Humane AI and the Stanford Institute for Human-Cantered Artificial Intelligence underscore this movement (Xu, 2019). HCAI, though not strictly defined, is primarly viewed as an AI development approach that promotes human welfare, transparency, and user control over data and algorithms. Shneiderman (2020) likened HCAI's philosophy to a contemporary Copernican shift, positioning humans at the heart of AI design. When reflecting on how AI tools like AI-Enhanced Education (AIEd) are applied within academic contexts, the term 'Digital Humanism' seems appropriate (Wiener Manifest Für Digitalen Humanismus, 2019). Schmölz (2020) traced the philosophical origins of digital humanism, observing a shift in the understanding of the Human Condition "Conditio Humana." While previously seen as a hallmark of rational thought during the Enlightenment, in this age of machine learning and AI, it is more about creativity and individual expression, freeing humans from purely calculative rationality. However, in recent years, we have seen Generative AI (GAI) encroaching rapidly onto what was previously considered as the sole domain for human beings.

Human-Centered AI and Academics

In the academic realm, AI tools present transformative opportunities for research methodologies (Jordan & Mitchell, 2015). These tools, ranging from neural networks to predictive analytics models, offer capabilities like data analysis and task automation (Kitchin, 2014; Brynjolfsson & Mitchell, 2017). Yet, there's a pressing need to address ethical concerns. Holmes et al. (2021) and Prunkl et al. (2021) emphasize the necessity for ethical frameworks and consideration of broader societal implications in AI research. Larsson (2020) further argues for a shift from mere principles to actionable processes in AI governance. A concerning facet of AI's impact on academia is the rise of 'paper mills' that produce potentially misleading or fraudulent papers. AI tools, such as ChatGPT, have exacerbated these challenges, necessitating meticulous scrutiny by academic journals (Barnett, 2023; Brainard, 2023; Liverpool, 2023). Governments (US, EU, UK, Australia), concerned by the repercussions of this technology, are stepping in to develop regulatory frameworks (EU: AI Act) to mitigate the possible drawbacks of the technology.

Regarding AIEd's integration into academia, it's pivotal for educators to possess a comprehensive understanding of AI to harness its potential to enhance educational outcomes (Crompton & Song, 2021). Baker and Smith (2019) identified three primary AI application perspectives: learner-facing, teacher-facing, and system-facing.

Learner-facing tools in AIEd offer personalized instructional strategies and feedback mechanisms catering to individual student needs (Sobel & Kushnir, 2006). These tools, fortified with gamification elements, not only boost student engagement but also foster critical skills such as problem-solving (Ghaban & Hendley, 2019) and lateral thinking. Examples include AI-powered writing aids that enhance student writing proficiency (Alharbi, 2023) and tools that help educators identify and support at-risk students (Topali et al., 2019).

From a teacher's perspective, the advent of AIEd necessitates the development of digital AI competencies to effectively harness AI tools (Niemi, 2021; Ng et al., 2023). These tools, including AI-driven educational platforms and assessment systems, can streamline teaching practices and elevate their efficacy (Akgun & Greenhow, 2021) and productivity Chounta et al., 2021). For instance, AI-Grader, an AI-powered evaluation tool, can reduce the assessment burden on educators and bolster their confidence in AI-driven educational technology (Nazaretsky et al., 2022) by serving as a preliminary tool before the instructor can allocate the final grades on specific assignments.

On the other hand, system-facing AIEd tools, which are less prevalent, enable data-sharing across educational institutions and cater to organizational tasks beyond individual educators or learners. Included productivity tools range from scheduling to predictive inspections, ensuring the seamless functioning of educational systems (Baker et al., 2019). Recent examples include AI-powered Google Workspace as well as Bing-powered Microsoft Office suite, which now consists of a wide range of tools and services that can enhance productivity and foster seamless information sharing (i.e., MS Co-Pilot).

Common Tools for Academic Research

In this section, we will review some of the tools from the three AI application perspectives defined by Baker and Smith (2019). In the "Learner-Facing" perspective, tools such as Consensus, Elicit, Inciteful, Laser, Litmaps, Rabbit, System Pro, Scite.ai, and Semantic Scholar are most common. These focus on the HCAI approach, emphasizing and prioritizing human education. *Consensus* is an AI-driven search engine that delivers evidence-based answers to scientific research, concisely presenting essential findings from peer-reviewed sources. *Elicit* employs language models to identify pertinent academic papers and supports diverse research activities, including integration with citation managers. *Inciteful* is an open-source tool that leverages citations to help users locate academic literature, offering unique tools such as the literature connector. *Laser* AI focuses on enhancing systematic reviews by offering semi-automated data extraction, which can significantly reduce review times. *Litmaps* produces visual maps of related articles, harnessing citation patterns to streamline the literature review process. Research *Rabbit*, a platform appreciated for its user-friendly interface, offers a simplified literature search and management process, complete with personalized recommendations. *System Pro* specializes in searching and analysing scientific literature, particularly health and life sciences, by merging large language models with structured data. *Scite.ai* aids in literature search by providing reliable answers from a vast research repository, summarizing content, and locating sources. *Semantic Scholar* is a comprehensive research tool powered by AI, with its Semantic Reader enhancing the reading experience by providing added context. In essence, these tools harness the power of AI to demystify complex academic literature, offering various features to assist researchers in diverse aspects of their work.

Similarly, from the teaching perspective, tools such as *Chat PDF* employ an AI-driven conversational interface to facilitate real-time understanding of academic papers, allowing users to pose questions directly to the document. *Explain Paper* streamlines the comprehension process by enabling users to upload papers, highlight portions, and receive explanations, making intricate content more digestible. *Lateral AI*, an app designed to streamline research, offers features like text search, organization of findings, and easy document viewing. *Open Read* promotes engagement with scholarly work by providing AI-crafted summaries and interactive papers, although it's still in its early stages of development. *Scholarcy* is an AI-powered tool that distills academic content into summaries, extracts structured data, facilitates collaborative note-taking, and even offers a browser extension for ease of use. *SciSpace Copilot* is a comprehensive research platform that automates repetitive tasks and provides access to an extensive metadata collection of papers. Lastly, based on the GPT-4 model, *Unriddle* deciphers intricate topics by summarizing them, allowing for interactive Q&A sessions, and offering a custom AI creation from any document, catering to a broad audience, from students to professionals. These and other similar tools cater to the diverse needs of academic writers, from content generation and editing to paraphrasing and compliance checking, facilitating a more streamlined and refined writing process.

Lastly, the most common tools currently developed to support deliverables utilize instruct- or chat-based LLM technology that can restructure and enhance human output. Tools such as *Jenni.ai* offer an array of functionalities like AI autocompletion, in-text citations, and paraphrasing using a blend of in-house AI systems, including GPT4 and ChatGPT3.5 Turbo. It supports multilingual content generation and translation, and boasts a built-in plagiarism checker. *Paper Pal* focuses on refining academic texts, leveraging AI to enhance clarity, coherence, and alignment with academic standards, while verifying technical and language compliance per journal standards. *Quillbot,* an AI-enhanced writing tool, provides grammar-checking, paraphrasing, and summarization capabilities, especially assisting non-native English speakers in refining their articulation. *Trinka*, tailored for academic and technical writing, helps researchers evelate the clarity and alignment of their content with academic norms. *Wisio.app*, an AI-infused platform, streamlines scientific writing by offering personalized text suggestions, citation extraction, translation, and English rectification tools, with plans to introduce more features in the future. Lastly, *Writeful* mirrors the offerings of *Paper Pal*, providing AI-driven editing for academic texts and offering language feedback, particularly beneficial for non-native English speakers.

Additionally, in the last six months, composite tools powered by LLMs and large foundational models have been made available to the research community. This development has been hastened by the release of PaLM2 and LlaMA2, which function as the core LLM in commercial and opensource tools (i.e., BabyGPT, Agent-GPT, Open-Interpreter, etc.) and have been used to train and fine tune specific large language models designed to excel in academics and research-based tasks. These include but are not limited to StarCoder, Cerebras-GPT GPT-NeoX, Polyglot, H2O.ai's h2ogpt, InternLM, Research Agent, Mosaic ML's MPT, Falcon LLM, and, notably, our previous iteration of TAUCHI-GPT. In combination with recursive logic platforms (i.e., LangChain Model and HuggingFace repository) as well as Agent-Module-Chains (AMC) using structured and inexpensive cloud compute (TextGen WebU, RunPod, etc.), creating customized pre-trained transformers, fine-tuned for specific tasks is the natural progression in academics and research.

DEVELOPING TAUCHI-GPT_V2 OFFLINE FINETUNED AGENT-BASED SYSTEM USING LLAMA2 13B

Building on tools such as LangChain and HuggingFace, as well as the plethora of fin-tuned pre-trained transformers, we developed TAUCHI-GPT_V1 and TAUCHI-GPT_V2. The former version (Farooq et al., 2023) was built on the forked open-source platform of Auto-GPT agent-based reductive architecture. By simply identifying specific goals and objectives, users could monitor and supervise autonomous agents solving complex problems using reductive reasoning and the underlying foundational model of choice (Chat-GPT, PaLM2, LlaMA2, etc.). However, in most cases, the computing necessary to ensure reliable execution of the foundational model required cloud resources and, therefore, a networked topology. Nevertheless, the results of the previous study (Farooq et al., 2023) showed that reductive

reasoning and autonomous agents powered by GPT4 could be very useful in solving complex research tasks. Therefore, building on previous research, we developed a localized version of the TAUCHI-GPT, which utilized a similar agent-based architecture built on the LangChain model but was powered by a fine-tuned version of LlaMA2 13B.

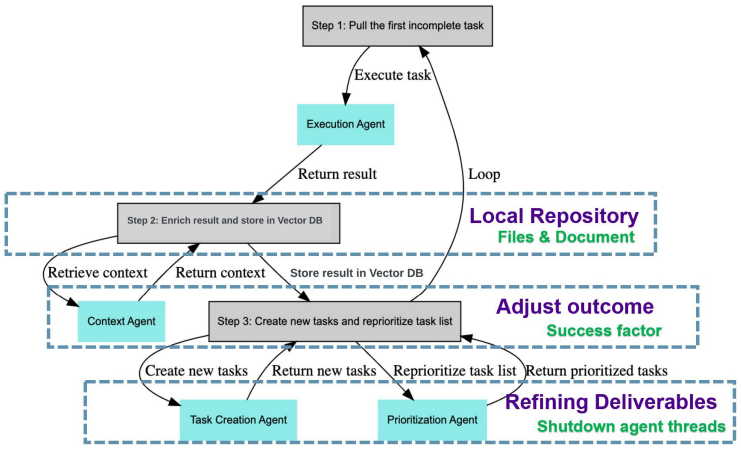

Figure 1: Design of the agent recursive "*Observe, Orient, Decide, Act*" (OODA) loop used in TAUCHI-GPT_V2.

LlaMA2 13-billion parameter version was finetuned using Gradient Development Platform. The gradient is a finetuning and inference tool that can be utilized for customizing Large Language Models (i.e., embeddings/vectorDB). Using Gradient API, it was possible to tune and get completions without developing any local system infrastructure. Additionally, we utilized Google Collab, a cloud-based platform for data analysis and machine learning research.

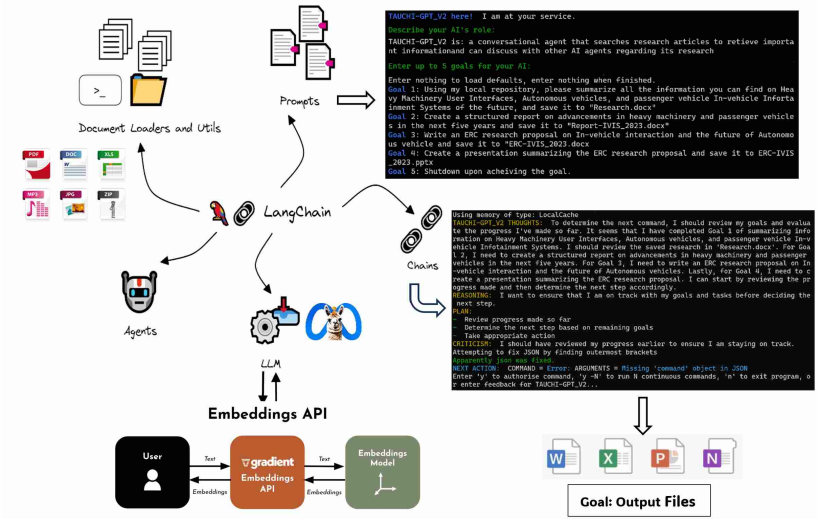

Figure 2: Design and topology of the TAUCHI-GPT_V2 offline localized system using the LangChain model and gradient embedding API.

The system ensured that we could execute Python code in a Jupyter Notebook environment to create and share computation files during development without needing local infrastructure (i.e., by using Google services and Google Drive). In combination with Google Collab and Gradient, we built TAUCHI-GPT_V2, an offline customized private SOC2 Type-1 compliant AI tool, finetuned on recent opensource research publications in software engineering (SE) and human-computer interaction (HCI).

Local System Topology

TAUCHI-GPT_V2 was designed to extend the capabilities of the linked LLM (LlaMA2) architecture by integrating supplementary modules and features to facilitate the generation of multimodal text. This system was designed to combine the local vectorDB and generate embeddings from user local repositories. These repositories included various document formats (docx, xlsx, pdf and other file formats) while participants could easily maintain within their local repository and manage them in real time. This ensured that users could easily interact with their local documents and extrapolate offline resources from TAUCHI-GPT_V2 similar to any remote LLM powered system (i.e., BARD, Bing, GPT4 etc.). Additionally, the fine-tuned "Code LlaMA 13b" resided locally on the system machine therefore, no internet connection was needed to run the system and no personal data was forwarded to the cloud. Thus, this architecture ensured a private and completely local environment for the users to conduct research and academic tasks utilizing their offline documents / knowledgebase.

USER STUDY & DATA COLLECTION

The user study was designed to understand how researchers at Tampere University would utilize TAUCHI-GPT_V2 in their routine daily tasks. The study included 18 participants over a period of 3 weeks. Traditionally, researchers utilize online resource services and search engines with optimised keywords to locate and incorporate the necessary knowledge/tools needed to complete their daily activities. The workflow meant that researchers might acquire knowledge while a task was being carried out, and we wanted to design our study to ensure that this process was not altered. Therefore, instead of using TAUCHI-GPT_V2 as a blackbox (i.e., *"Chat-GPT", "Code Interpreter," or "Open Interpreter")* to complete their tasks from start to finish, we wanted the participants to use TAUCHI-GPT_V2 as a centralized tool, which would assist the researchers in optimizing their workflow and acquiring the knowledge necessary for boosting their productivity.

Eighteen university researchers from the Department of Computer Science were asked to list their commonly performed task categories before the experiment. As all the participants were from the computer science faculty most of their daily task categories were similar. However, the experimenter reviewed the category list in advance and ensured any divergent overly complex task type was removed from the list, which may not be possible to complete using TAUCHI-GPT_V2. The list of three categories of tasks, highlighted below, was used as a basis for the user study, where each task required roughly

15 minutes to complete successfully. Each participant created seven custom tasks from these three categories relevant to their research area and project focus totalling 21 tasks (Table 1). Additionally, participants created another 21 tasks similar to the original list to serve as a control for the experiment for each of the three categories: "*Editing / Writing Summarizing*," "*Troubleshooting / Writing Python code*," "*Brainstorming ideas for experimental design*" (42 in total).

The experimenter then randomly assigned one of the two sets of identical 21 tasks to either the TAUCHI-GPT_V2 or the Common Search Engine list. These task lists (one for each system) were researchers' TODO lists for the next three weeks. The participants were free to randomly select any task on the TODO list as long as they ensured both similar tasks were completed before moving down their TODO list. This meant that once a task was selected from one TODO list (i.e., TAUCHI-GPT_V2), the participant needed to complete a similar task from the second TODO list (Custom Search Engine). The participants were also instructed to alternate their TODO lists, which meant they needed to select one unique task from one list, select the following unique tasks from the second TODO list, and so on.

In the case of conventional search engines and online services, participants were free to use the SE of their choice (Google, Bing, Github, PythonLib, etc.) without any LLM or AI tools. For the TAUCHI-GPT_V2, each participant first optimized their version of the tool. It was achieved by creating an offline workspace for each participant, including all relevant documents, reports, publications, research books/textbooks, etc. All their documents were ingested into the TAUCHI-GPT_V2 to create a vectorDB effectively finetuning the system for each user. Furthermore, participants were also allowed to add any documents they needed during the 3-week user study.

Once the tools were set up on each participant's machine, they were given four pilot tasks to get familiar with the system. The system included a timediary, which the participants filled out as they went through the task list. This included the start and end time of each task and any issues they may have faced while carrying out the tasks. Additionally, other objective measurements collected by the system included logging Task Completion Times and user inputs for both the search engine queries, the prompts used in the TAUCHI-GPT_V2, and the idle time on the user's machine between tasks/user inputs.

Subjective parameters were also collected automatically by the system through automatic user forms generated after the completion of each task. These included "Number of Errors" (*No. Errors*), "Numbers of Prompt Needed" (*No. Prompts*), "How satisfied were they of the outcome" (*Satisfaction Scale x/7*), and "Overall efficiency of the outcome" (*Efficiency Scale x/7*). The purpose was to collect information regarding user satisfaction of the results generated by each system and how efficiently the participants could complete the task.

Table 1. The three categories of tasks as well as the type of tasks used in the TODO list.

Editing / Writing Summarizing Text	Troubleshooting / Writing Python code	Brainstorming ideas for experimental design
Creating plans to improve writing process	Creating Methodology (high level interface)	Creating Overall design / plans
Improving text language	Researching functions / packages / APIs / modules	creating / understanding User I/Os
Summarizing text	Understanding syntax (what does a piece of code do)	Developing Hypothesis for user testing
Reviewing text	optimizing / troubleshooting code (syntax & semantics)	find existing research & testing methodologies
Critiquing text	Creating an Effective deployment strategy (HW/Platform)	Creating scenarios / goals for various tasks and hypotheses
Comparing different text	Comparing code base	Projected outcome / results
Combining main points from multiple text	Merging projects	Identifying optimum testing tools

RESULTS & DISCUSSION

As discussed, to evaluate the two approaches, we utilized four metrics: 1) Task Completion Times (TCTs in Fig. 3) were automatically recorded through the usage diary, 2) "Number of Prompt used" (#PU) for TAUCHI-GPT_V2, or the "Number of Links used" (#LU) for conventional search engines and MS office suite/IDE were also recorded using the participants usage logs (Fig. 3 and Fig. 4) "Results Satisfaction Score" (RSS in Fig. 5), and 4) "Overall Efficiency Score" (OES in Fig. 6) from 0–7 on a Likert scale after each completed task.

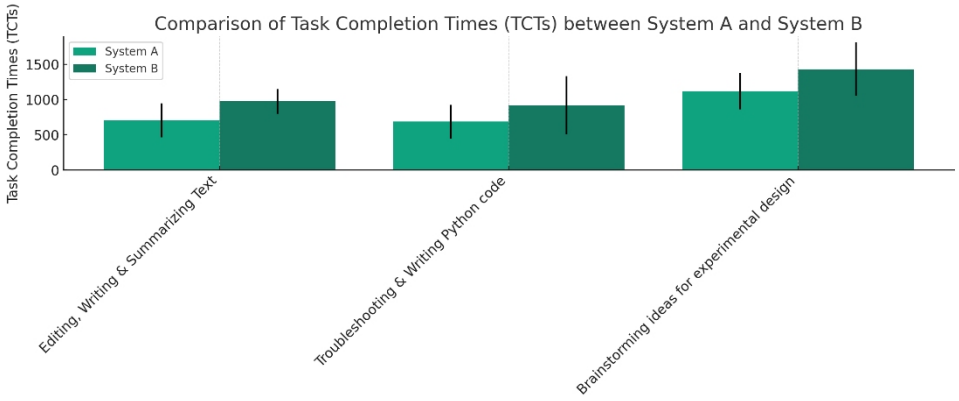

Figure 3: Comparing TCTs between System A (TAUCHI-GPT_V2) and System B (MS office suite and search engine of preference) across the three task categories, *with the bars representing the mean values and the error bars representing the standard deviations.*

Task Completion Times for "Editing, Writing, & Summarizing Text" resulted in a p-value of 0.036, suggesting a statistically significant difference between System A (TAUCHI-GPT_V2) and System B (Conventional MS Suite and preferred Search Engine), with System A having a lower mean time as shown in Fig. 3. For "Troubleshooting & Writing Python code," the p-value was 0.222, indicating no statistically significant difference in TCTs. Similarly, for Category 3, "Brainstorming ideas for experimental design," the p-value was 0.121, showing no statistically significant difference in task completion times between the two systems.

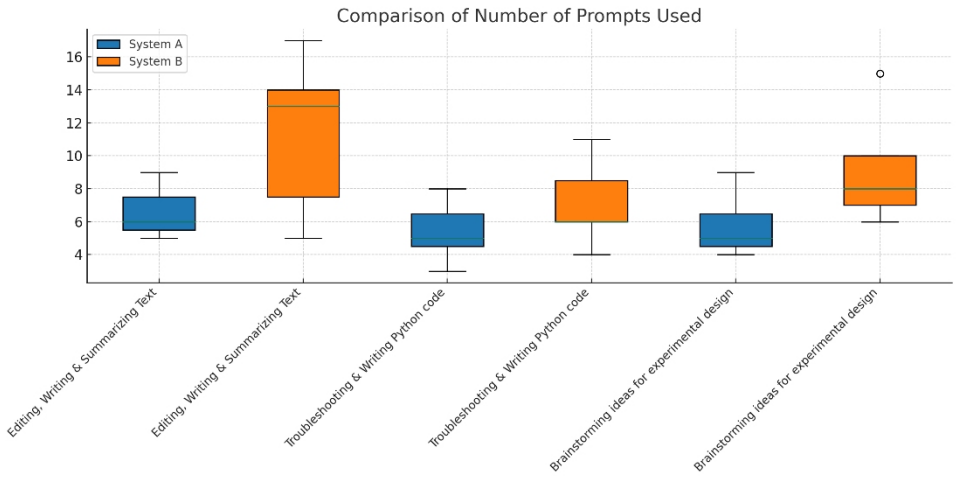

Figure 4: Comparing the number of prompt/links used to complete each task for the three categories between System A (TAUCHI-GPT_V2) and System B (MS office suite and search engine of preference). Each pair of boxes represents the interquartile range (IQR), the line inside the box is the median, and the whiskers represent the range within 1.5 * IQR.

Comparing the number of "Prompt / links" used to complete each task, we observed a difference between System A and System B. The p-value for "Editing, Writing, & Summarizing Text" was 0.024, suggesting a statistically significant difference in the number of prompts/links used between System A (TAUCHI-GPT_V2) and System B (MS Office and Search Engine), with System A having a lower mean number of prompts used. This correlated with the TCT results. Additionally, for "Brainstorming ideas for experimental design," the p-value was 0.048, suggesting a statistically significant difference in the number of prompts used between the two systems, with System A having a lower mean number of prompts used. Whereas results from "Troubleshooting & Writing Python code" did not show a meaningful difference between the two systems.

We recorded similar trends by analysing the Results Satisfaction Score (RSS) and the Overall Efficiency Score (OES). For "Editing, Writing & Summarizing Text," System A (TAUCHI-GPT_V2) had significantly higher RSS and OES compared to System B (MS Office and SE of Choice) with p-values below 0.05, indicating a statistically significant difference. Moreover, for

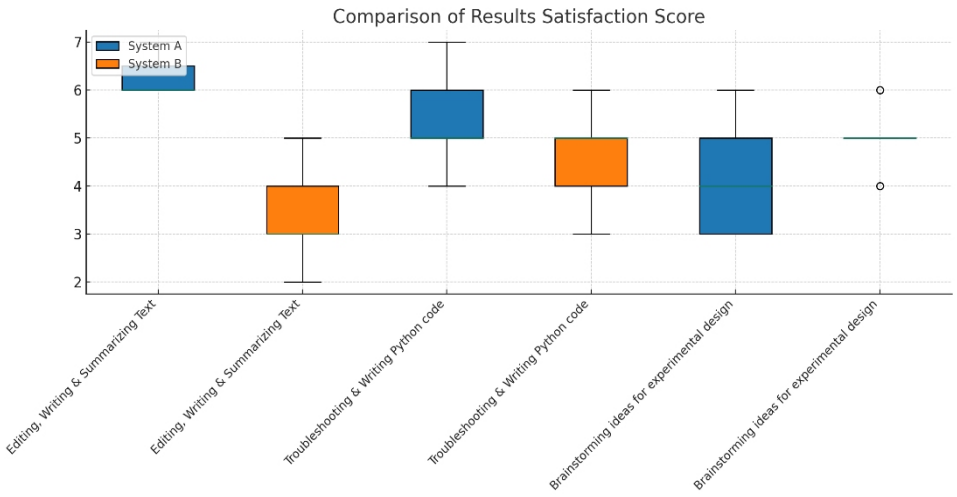

Figure 5: Comparing the results satisfaction score (RSS) between System A (TAUCHI-GPT_V2) and System B (MS office suite and search engine of preference). Each pair of boxes represents the interquartile range (IQR), the line inside the box is the median, and the whiskers represent the range within 1.5 * IQR.

"Troubleshooting & Writing Python code," System A had higher RSS and OES values than System B. Still, the difference in RSS is not statistically significant (p-value = 0.126). In contrast, the difference in OES is statistically significant (p-value = 0.00675). Conversely, if we look at the results for category 3, "Brainstorming ideas for experimental design," System B (MS Office Suite and SE of Choice) has higher RSS and OES scores compared to System A (TAUCHI-GPT_V2), with the differences being statistically significant (p-value = 0.020 for OES). This means that participants preferred using their current tools and workflow for tasks pertaining to this category of planned and structured user studies as they trusted their current approach more than the new system.

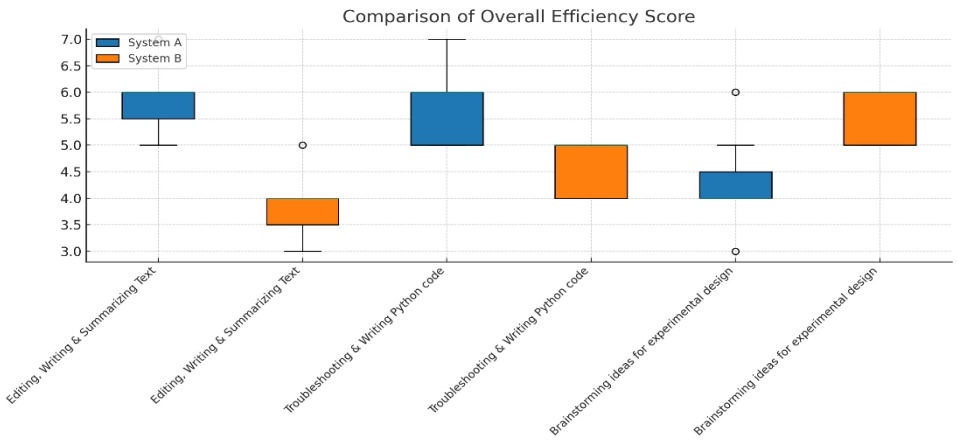

Figure 6: Comparing overall efficiency score between System A (TAUCHI-GPT_V2) and System B (MS office suite and search engine of preference). Each pair of boxes represents the interquartile range (IQR), the line inside the box is the median, and the whiskers represent the range within 1.5 * IQR.

CONCLUSION

With the advent of generative AI tools and Large Language Models (LLMs) like ChatGPT, Llama2, and Claude, producing distinctive content is now easier than ever. However, most of this development is happening outside the sphere of research and academics. This paper seeks to bridge the gap by introducing TAUCHI-GPT_V2, an evolved academic research assistant built upon the foundations of previous autonomous agents and open-source LLMs, using the LangChain architecture and leveraging LLaMA2-13b as its core LLM. TAUCHI-GPT_V2, conceptualized as an open-source initiative, ingests user-specific data, offering contextually pertinent results through a sophisticated, user-friendly chat interface, employing a custom offline localized vectorDB for nuanced parsing user's personal files. The model's practicality was assessed through its integration into the daily workflows of academic researchers, with preliminary findings indicating its substantial efficacy and reliability in enhancing user workflow for academic research tasks, including citation and information sourcing, and minimizing hallucination. Multiple metrics were used to analyse the data, and we found a notable variance between System A (TAUCHI-GPT_V2) and System B (Conventional MS Suite and preferred Search Engine) concerning task completion times (TCTs), usage, "Results Satisfaction Score" (R.S.S), and "Overall Efficiency Score" (OES). Particularly in tasks associated with "Editing, Writing, & Summarizing Text," System A depicted statistically significant efficiency.

Although further investigation is necessary to explore and qualify the use cases and effectiveness of TAUCHI-GPT_V2, the present study demonstrates the feasibility of integrating offline localized AI tools with fine-tuned LLMs for R&D tasks into a central interface that can enhance productivity. Yet, more comprehensive research is needed to delve into the specific contexts and domains where TAUCHI-GPT_V2 could deliver the greatest benefits. With rapid development in this area, such as Microsoft's AutoGEN (2023) project as well as Google's Projects like Tailwind, NotebookLM (2023), or DuetAI workspace (2023), and Meta AI and EMU (2023), it is crucial to develop and understand the role of offline localized open-source models similar to TAUCHI-GPT_V2 that are targeted towards academic research. Additionally, our future work will also focus on leveraging these open-source tools as a GUI implementation that can be utilized on various platforms including mobile devices. This approach will involve using the LangChain model and offline document repositories to fine-tune responses tailored to specific research workflows for users on mobile devices (Android and iOS).

ACKNOWLEDGMENT

This work was funded by Business Finland as part of the AMICI project and research was conducted in collaboration with Bentley University, in Massachusetts, USA.

REFERENCES

AI Act: a step closer to the first rules on Artificial Intelligence (May 2023). Human-centric and ethical development of Artificial Intelligence (AI) in Europe. EU parliament website https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

Akgun, S., & Greenhow, C. (2021). Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings. Ai and Ethics. https://doi.org/10.1007/s43681-021-00096-7

Alharbi, W. (2023). AI in the Foreign Language Classroom: A Pedagogical Overview of Automated Writing Assistance Tools. Education Research International. https://doi.org/10.1155/2023/4253331

AutoGen: Enabling next-generation large language model applications. Microsoft Opensource Project. https://www.microsoft.com/en-us/research/blog/autogen-enabling-next-generation-large-language-model-applications/

Baker, T., Smith with Nandra Anissa, L., Sheehan, K., Ward, K., Waters, A., Berditchevskaia, A., Van Den Berg, C., Campbell, N., Candsell, O., Casasbuenas, J., Cinnamon, J., Copeland, E., Duffy, E., Hannon, C., John, J., Grant, J., Klinger, J., Latham, M., Macken, C.,... Ward-Dyer, G. (2019). Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges. www.nesta.org.uk

Barnett, A. (2023, May 31). Scientific fraud is rising, and automated systems won't stop it. We need research detectives. The Conversation. https://theconversation.com/scientific-fraud-is-rising-and-automated-systems-wont-stop-it-we-need-research-detectives-206235?utm_source=substack&utm_medium=email

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K. L., Steffens, N., Wiles, J., & Worthy, P. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. Computers in Human Behavior, 141, 107617. https://doi.org/10.1016/J.~CHB.2022.107617

Brainard, J. (2023). New tools show promise for tackling paper mills. Science (New York, N. Y.), 380(6645), 568–569. https://doi.org/10.1126/science.adi6513

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358, 1530–1534. https://doi.org/10.1126/science.aap8062

Chounta, I.-A., Bardone, E., Raudsep, A., & Pedaste, M. (2021). Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support Their Practice in Estonian K-12 Education. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-021-00243-5

Crompton, H., & Song, D. (2021). The Potential of Artificial Intelligence in Higher Education. Revista Virtual Universidad Católica Del Norte. https://doi.org/10.35575/rvucn.n62a1

DuetAI: Workspace with AI tools in the clould https://cloud.google.com/duet-ai

Farooq A., Raisamo R., Kangas J. (2023) TAUCHI-GPT: Leveraging GPT-4 and Auto-GPT to create a Multimodal Open Source AI Research tool. In proceedings of 14th AHFE International Conference on Applied Human Factors and Ergonomics: Artificial Intelligence and Social Computing, Huwaii, USA 4–6 Dec 2023.

Ghaban, W., & Hendley, R. J. (2019). How Different Personalities Benefit From Gamification. Interacting With Computers. https://doi.org/10.1093/iwc/iwz009

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in Education: Towards a Community-Wide Framework. International Journal of Artificial Intelligence in Education, 32(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. Asian Journal of Law and Society, 7(3), 437–451. https://doi.org/10.1017/als.2020.19

Liverpool, L. (2023). AI intensifies fight against 'paper mills' that churn out fake research. Nature, 618(7964), 222–223.

MetaAI (2023): The Llama Ecosystem: Past, Present, and Future https://ai.meta.com/.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' Trust In AI-powered Educational Technology and a Professional Development Program to Improve It. British Journal of Educational Technology. https://doi.org/10.1111/bjet.13232

Niemi, H. (2021). AI in Learning. Journal of Pacific Rim Psychology. https://doi.org/10.1177/18344909211038105

Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI Digital Competencies and Twenty-First Century Skills in the Post-Pandemic World. Educational Technology Research and Development. https://doi.org/10.1007/s11423-023-10203-6

NotebookLM: Extension of the Tailwind project. an AI notebook for everyone https://blog.google/technology/ai/notebooklm-google-ai/.

Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. Nature Machine Intelligence, 3(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Schmölz, A. (2020). Die Conditio Humana im digitalen Zeitalter. Medien Pädagogik: Zeitschrift Für Theorie Und Praxis Der Medienbildung, 208–234. https://doi.org/10.21240/mpaed/00/2020.11.13.x

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. AIS Transactions on Human-Computer Interaction, 109–124. https://doi.org/10.17705/1thci.00131

Sobel, D. M., & Kushnir, T. (2006). The Importance of Decision Making in Causal Learning From Interventions. Memory & Cognition. https://doi.org/10.3758/bf03193418

Topali, P., Ortega-Arranz, A., Dimitriadis, Y., Martínez-Monés, A., Villagrá-Sobrino, S., & Asensio-Pérez, J. I. (2019). "Error 404- Struggling Learners Not Found" Exploring the Behavior of MOOC Learners. https://doi.org/10.1007/978-3-030-29736-7_56

Weidener, L., & Fischer, M. (2023). Artificial Intelligence Teaching as Part of Medical Education: Qualitative Analysis of Expert Interviews. Jmir Medical Education. https://doi.org/10.2196/46428

Wiener Manifest für digitalen Humanismus. (2019).

Xu, W. (2019). Toward Human-Centered AI. Interactions. https://doi.org/10.1145/3328485