UNIVERSITY
OF WOLLONGONG
AUSTRALIA

HDR HELPFUL HINTS

From the Dean of Graduate Research, Simon Moss.

Technologies and apps that enhance the benefit of AI

SUMMARY

To enhance the power of generative AI tools, such as Chat GPT, companies have developed an array of other technologies. These technologies include

- retrieval augmented generation—to enable AI tools to extract data and information from many sources while completing tasks,
- AI autonomous agents—that can perform a series of tasks autonomously, without the need for users to intervene,
- AI devices—devices besides computers, tablets, and phones to apply AI, and
- many other advances.

RAG OR RETRIEVAL AUGMENTED GENERATION

Context windows

Often, you want to enter a significant amount of information into your prompt. To illustrate

- you might enter a prompt like "can you summarise the following documents",
- and then you might copy and paste many documents into the prompt.

Some tools enable you to upload these documents or data separately rather than copying this information into the prompt. Nevertheless, the amount of information and data you can enter into a single prompt, called the context window, are limited. If you exceed this limit, the AI tool will tend to disregard the additional data. Admittedly, in some tools, this context window or limit is large—sometimes approaching about 700 000 words or an hour of video. But even this limit can impede some tasks.

Introduction to RAG or Retrieval Augmented Generation

To overcome these limitations, users can apply an approach called RAG. In essence, RAG enables AI tools to derive information from multiple sources, such as datasets and websites. The benefit of RAG is that

- AI tools can utilise many data sources to generate answers that would otherwise surpass the context window,
- because the AI tools can utilise multiple data sources to inform answers, the likelihood of errors or hallucinations diminish,
- users often receive more explicit information about which sources of data were utilised to generate the answers.

The following example outlines a tool that applies RAG.

Verba
Functions • A chatbot that enables users to ask questions about multiple documents, in diverse formats. • In essence, Verba is an interface to help users generate answers from multiple sources of data. • When users are a question, such as "What are the risks of AI", the tool retrieves all data sources that could be relevant to this question—and then uses an AI model to generate an answer from these data sources. • You can use any open-source AI models you choose—rather than, for example, only GPT4o—to analyse the data. • In addition, you can decide which documents or datasets you want to upload—such as the medical records or databases you want to analyse.

Occasionally, the RAG does not extract the data or information that is most relevant to the task. Partly to redress these concerns,

- a tool, called RAG Workbench, helps developers evaluate, debug, and enhance RAG applications,
- the tool monitors the retrieval and use of data to check the application is operating as planned.

AI AGENTS

AI agents—such as Agent GPT, CognoSys, Devin, Devika, Godmode, and Younet—are tools that may complete a series of tasks autonomously while you complete other activities. For example, you could simply enter a command, such as "create a website about a sock business", "organise a trip for me to Paris", and so forth. The agent would then

- plan which tasks need to be completed,
- execute these tasks without your intervention.

Therefore, in contrast to usual chatbots, like Chat GPT, you would not need to interact with the agent, besides the original command or specific questions you may receive. The agent would achieve your goal while you complete other activities. Agents are limited now but are gradually developing. As they become more efficient and independent, these AI agents will be referred to as AI autonomous agents.

AUTO-GPT

Functions

- When users specify a goal, such as "organise an online conference", the tool will divide this goal into sub-tasks and then attempt to complete the sub-tasks.
- Specifically, users merely need to specify the name, role, and objective of the agent. Next, users will specify up to five ideas on how the agent can achieve this objective.
- Auto-GPT will then attempt to achieve this objective independently. The user will not have to prompt each step.
- Auto-GPT can utilise information from the internet to complete these tasks.

Access

- Users must be able to access a paid OpenAi account to receive a code, called an API key.
- Next, they access Auto-GPT from GitHub.
- Finally, they must install this program in a development environment, such as Docker. A development environment is a workspace in which users can change software without disrupting the live environment.

Potential applications

- Auto-GPT can be useful in business to create business plans, review which products to purchase, conduct market research, and so forth.

Limitations.

- Auto-GPT may be more likely to commit errors and present false information than other AI tools—because users do not intervene to correct mistakes.
- Auto-GPT needs to utilise the Open AI API continually and thus may be costly to operate.
- Auto-GPT sometimes performs the same tasks repeatedly because the context window is finite. S, the agent may have forgotten that a task has been completed.

BABY AGI

Functions

- Designed to manage a wide range of tasks, gradually adapting from past experiences like a baby learning to navigate the world.
- Unlike most other AI tools, Baby AGI can decide which of multiple tasks to complete, based on previous experience, and therefore can perform somewhat autonomously.
- The tool is limited now but will become gradually more useful in the future.

SUPERAGENT

Functions

- Enables users with no coding skills to create AI agents

Multi-agent systems

To illustrate, some organisations, such as Microsoft, are developing multi-agent systems, in which several distinct AI agents collaborate to complete multifaceted tasks. The benefits of multiple agents, collaborating in parallel, are that

- such AI applicants complete many tasks more rapidly, because each agent will specialise in a particular capability,
- developers can gradually enhance the capability of these multi-agent applications over time,
- these systems are not as vulnerable to problems because they do not depend on a single agent or tool that could be faulty,

- the collective intelligence of multiple agents could uncover remarkable innovations that no single agent could generate.

To illustrate, Microsoft have released AutoGen Studio—a tool that enables anyone, with minimal experience in coding, to develop these multiple-agent applications. Users can select a series of pre-existing agents or can develop their own agents. They can then customise each of these agents.

AI DEVICES

Typically, to utilise generative AI, most people use their computer, tablet, or phone. However, individuals can now purchase a range of other devices that can facilitate interactions with AI tools. Here are a couple of examples:

DEVICE	DETAILS
Ai Pin	**Functions** • A wearable AI device. You can attach the device to your shirt, for example. • You can simply ask the device to remember some video or audio information. • Or you can ask the device questions about the environment or about the world, similar to Siri. • This device is designed to be more seamless to use than a mobile—and thus is not as likely to distract your attention. **Costs and concerns** • Some users feel the device answers too slowly, however. • The devices now cost around $700 US.
Copilot+ PCs	**Functions** • Computer laptops designed to work effectively with AI—such as generate fast AI images using Cocreator or to translate speech to English. • As an excellent illustration, some PCs may store all the pages—such as websites—you observe during the day—using a feature called Recall. Then, you can ask questions about these pages, such as "Show me the page in which I learned about watches". • Therefore, this function may enable you to review or locate any materials you watched during the day. **Concerns** • One concern is that, at this time, Recall stores data in plain text on your device. Hackers and Infostealer trojans may be able to access the information in this text, such as account numbers and other private data.
iOS18	• Apple have now launched IOS18. • This operating system on iPhones will now include many AI features, called Apple Intelligence.

	• These features will enable automated emails, improved Siri capabilities, and many other benefits. Apple intelligence will include • an AI writing assistant that can help you write, change your tone, and generate arguments in Notes, Mail, and other apps, • outline the preceding emails to a thread—so you do not have to reread the thread, • delete distracting features in photos, • verbalise instructions, such as "play this podcast", and the device will activate the right app. Apple intelligence will most likely utilise Siri, Chat GPT, and Gemini to complete AI tasks on Apple devices, such as iPhone 16.
Rabbit R1 	**Functions** • You can speak to this AI tool, and the tool will respond, such as call an Uber. • The tool includes a camera so, for example, you can use Rabbit R1 to name various objects in the environment. • The device will generate images. **Costs and concerns** • This device is, arguably, no more useful than a mobile phone yet but will develop more features over time. • Costs a few hundred dollars.
Meta Ray-Ban smart glasses 	• They resemble normal glasses but include speakers, microphones, and a camera. • An AI can respond to what you see, such as name the objects.

AI ACCELERATORS

An AI accelerator is a class of computer hardware that is designed to facilitate AI. These accelerators are typically electronic circuits on microchips that can perform many calculations in parallel—and often found in mobiles, computers, servers, and other devices.

History

In most computer systems is the CPU, or central processing unit. The CPU is a digital circuit or processor that executes key instructions in a computer program, such as arithmetic or logic. Typically, digital circuits or co-processors complement the CPU and specialise in other tasks. For example, video cards and graphics processing units improve the graphics or digital images on computer screens. Some of these co-processors, including graphics processing units, perform multiple operations in parallel—and, therefore, were later adapted to complete very complicated tasks, including AI and neural networks.

Features of AI accelerators

The goal of AI accelerators is to apply neural networks, and thus support AI, as rapidly and as efficiently as possible—to minimise delays and energy consumption. To achieve this goal, AI accelerators can

- conduct multiple calculations in parallel,
- optimise the use of memory
- conduct arithmetic at low precision—in essence, with no more decimal places than needed.

AI DATASETS

Users often want to develop and train their own AI tools. To achieve this goal, users need access to significant datasets—such as social media posts, academic articles, and anything else that could be useful to the tool. Consequently, companies are now enabling users to access huge pools of data. One example is DCLM-POOl. Here is some information about this pool of datasets:

- comprises 240 trillion tokens, extracted from Common Crawl, and is thus massive,
- supplies free software to process large datasets,
- provides training protocols you can use to develop an AI tool.

Indeed, hundreds of datasets are available. This page outlines many of the public datasets that individuals can use to training AI and machine learning in specific domains.

AI APPS

Chat GPT desktop app

Initially, to use Chat GPT or other generative AI tools, users would visit the relevant websites. Now, these tools are often available as desktop apps. For example, you can download the Chat GPT desktop app. The app offers several benefits over the website. For example

- the app will conduct data analysis more effectively, such as clean datasets and create charts or tables,
- you can then customise or refine these charts,
- you can then click specific rows, columns, or cells in these tables to explore specific findings in more depth,
- you can use voice mode, in which you can voice your prompt rather than write the text,
- you can more readily upload files from Microsoft OneDrive or Google drive—such as Excel, Word, or Powerpoint files.

HYBRID APPROACHES

Besides technologies and apps, individuals have identified novel approaches or practices to enhance the utility of AI. For example, many organisations are now developing hybrid approaches—that is, procedures that integrate AI with humans. One example is Jung.ai: an approach that blends AI with the expertise of psychologists. This approach is designed to diminish the costs of psychology services as well as circumvent long waiting lists.

Jung.ai
Procedure • an AI tool first assesses the mental health of users and generates a report, • a human psychologist then assesses this report, • the human psychologist then meets the user online to discuss a treatment plan, • the AI tool then helps the user follow this treatment plan, • at regular times, the human psychologist will review and monitor progress as well as adjust the treatment plan when needed.

Challenges

These hybrid models are obviously more efficient than practices that depend solely on humans. Nevertheless, these hybrid models may generate some complications:

- whether consumers are as satisfied with these services obviously warrants further research,
- the practitioners or employees themselves sometimes perceive their role as less meaningful, diminishing their intrinsic motivation (e.g., Mehler & Krautter, 2024); indeed, the diminished effort these individuals experience may limit the sense of meaning and purpose they attach to their role.