ChatGPT “Deep Research” vs. GPT-4 vs. GPT-3.5: Feature Comparison

ChatGPT’s “Deep Research” is a new advanced mode (introduced in 2025) that differs significantly from earlier GPT-4 and GPT-3.5 models in capability and workflow. Below is a detailed comparison focusing on key areas:

Integration with Search Tools and Real-Time Web Access

- **ChatGPT Deep Research:** Fully integrates with web search and browsing. It can autonomously query search engines, visit dozens or even hundreds of webpages, and retrieve up-to-date information in real time. The model reads and analyzes online texts (including articles, PDFs, and images) and pivots its search strategy based on what it finds, acting as an agent that *“finds, analyzes, and synthesizes hundreds of online sources”* into its answer. This allows it to handle current events and niche queries that go beyond its training data.

- **GPT-4 (standard):** No built-in live web access in its default ChatGPT form. GPT-4’s knowledge is limited to its training cutoff (it cannot fetch new information from the internet on its own). While early experiments with a browsing plugin or Bing’s GPT-4-powered chat enabled some web queries, standard GPT-4 will not autonomously search the web for you. Any “real-time” info it provides is either from memory (which can be outdated) or from user-provided data. It relies on the user to supply relevant text or use external tools if up-to-date information is needed.

- **GPT-3.5 (standard):** No web integration at all. Like GPT-4, it only knows information up to its training cutoff and cannot perform any live searches. It’s even more constrained by an older knowledge base (often up to 2021) and will not retrieve new data. Any research beyond its knowledge must be done by the user and fed into the model manually.

Ability to Handle Complex Multi-Step Research Tasks

- **ChatGPT Deep Research:** Designed explicitly for complex, multi-step reasoning and research. It uses a specialized reasoning model (a version of OpenAI’s *o3* model) with chain-of-thought capabilities to break down hard tasks into smaller steps. Deep Research will **autonomously plan and execute a research strategy** for a given query: it may perform several rounds of searching, follow links, gather statistics, and refine the query based on intermediate findings. In effect, it behaves like a virtual research analyst, conducting *“multi-step research on the internet for complex tasks”* without needing step-by-step prompts for each sub-task. This allows it to answer highly involved questions (spanning multiple domains or requiring data synthesis) that earlier models would struggle with.

- **GPT-4 (standard):** Exhibits strong reasoning in a single turn, but **does not autonomously perform multi-step research**. GPT-4 can handle complex queries better than GPT-3.5 (thanks to a larger context and more advanced reasoning), and it can follow multi-step instructions *within one prompt* (for example, if you ask it to outline, then elaborate, it will try to do so internally). However, it won’t on its own decide to perform a sequence of searches or break the problem down unless the user explicitly guides it. Prolonged research tasks typically require the *user* to break the inquiry into parts and ask multiple questions in

sequence. Essentially, GPT-4 will give one-shot answers based on its existing knowledge; any deeper digging or iterative approach must be manually driven by the user or facilitated by outside tools.

- **GPT-3.5 (standard):** Much more limited in handling complex or multi-step tasks. GPT-3.5 often gives brief responses and can lose context over long dialogues. It doesn't self-organize a research plan or iterative thinking beyond what's in the prompt. For complicated questions, it may produce a superficial answer or get confused, requiring the user to break the task into simpler sub-questions. Any multi-step reasoning has to be coaxed explicitly, as it will not independently perform a long chain of reasoning or lookups. This makes it impractical for deep research without constant user intervention.

Structured Output Formatting (Tables, Reports, Summaries)

- **ChatGPT Deep Research:** Outputs are formatted as **comprehensive reports** rather than just short answers. Deep Research is built to deliver findings in an organized way – often with an introduction, multiple sections or headings, bullet lists, and even tables or charts for data. The model can compile comparison tables or structured bullet-point lists when appropriate (for example, comparing products or listing statistical findings side by side). The result is a polished, multi-page report that reads like a researched article, making it easy to digest complex information. This structured format is generated automatically based on the query (e.g. if asked for a comparison or overview, it might produce headings and tables without needing extra prompting for format). *Deep Research's reports can run for thousands of words* and are logically organized, something earlier models would rarely do unassisted.

- **GPT-4 (standard):** Generally provides well-written **paragraph-style answers** by default. It is capable of producing structured content (like lists, outlines, or tables) *if the user specifically requests it*, and it's better at this than GPT-3.5 due to its larger capacity and instruction-following. For instance, GPT-4 can create a table or an outline if prompted, and it can maintain format over a longer response. However, it *does not inherently output a full structured report* on its own initiative. The default interaction is Q&A; a user typically gets an answer in a few paragraphs or a short list. Long-form structured outputs (with multiple sections) from GPT-4 require careful prompting or breaking the task into parts. In summary, GPT-4's formatting is on-demand – it can do a decent job with it, but it isn't dedicated to report generation like Deep Research is.

- **GPT-3.5 (standard):** Outputs are usually shorter and more simplistic in structure. By default, GPT-3.5 gives a quick answer (often a single paragraph or a basic list). It *can* produce structured text (for example, bullet points or a simple outline) if asked, but it's more prone to formatting errors in very long outputs and can forget instructions in lengthy responses due to a smaller context window. GPT-3.5 would not spontaneously create multi-section reports with subheadings; it requires explicit step-by-step prompting to even approach that. Even then, the coherence and formatting consistency over a long report are not as reliable. In practice, GPT-3.5 is best for short answers or simple lists, not for complex structured documents.

Reliability, Accuracy, and Citation of Sources

- **ChatGPT Deep Research:** Emphasizes **accuracy and source citation** in its design. Every Deep Research report is *"fully documented, with clear citations"* for the information it presents. The model will reference its sources (often with footnote-style or inline citations)

for factual claims, allowing the user to verify each point. Because it pulls content directly from current web sources, it tends to stick to verifiable facts and quote or paraphrase them rather than invent facts. This significantly reduces blatant hallucinations, though it's not foolproof – Deep Research can still misinterpret data or assemble facts incorrectly (and OpenAI notes it *"may struggle with distinguishing authoritative information from rumors"*). In testing, it has produced detailed reports with dozens of citations (e.g. a 2,000-word biography citing 17 different sources). Its improved reasoning model also means it makes fewer logical errors or basic mistakes than GPT-3.5, and even fewer than standard GPT-4 in some domains. **However, it's not infallible** – early users found that it can miss subtle details or context (for example, misjudging timelines of a person's career from source data) and it will confidently present whatever information it found, so errors in sources can propagate into the answer. The key advantage is that you *can trace its statements back to references*, making verification or correction easier.

- **GPT-4 (standard):** Highly reliable in knowledge *within its training data*, but with notable caveats. GPT-4 significantly improved factual accuracy and reduces hallucinations compared to GPT-3.5, yet it **provides no source citations by default**. Any statement it makes cannot be directly verified unless the user checks externally. GPT-4 might occasionally **hallucinate** convincing-sounding facts or references, especially if pushed beyond its knowledge cutoff or asked for extremely detailed data. OpenAI has indicated that even the advanced research mode (Deep Research) can hallucinate at times, though at a lower rate than standard models. In normal usage, GPT-4's answers are only as accurate as its training; for recent events or very specific statistics, it may be outdated or unsure (often it will admit not knowing something after 2021). It does have a much stronger grasp and usually avoids obvious errors on well-covered topics, but subtle errors can still occur, and **the lack of citations means users must trust but verify** important facts on their own.

- **GPT-3.5 (standard): Less reliable and more prone to errors**. GPT-3.5 was the earlier ChatGPT model and is more likely to produce incorrect information confidently. It has a narrower knowledge base (typically up to 2021) and tends to guess or fill in gaps if it doesn't know something, which can lead to fabricated facts. It also cannot cite sources in a meaningful way – if asked for references, it might either refuse (saying it cannot browse) or it could even invent plausible-looking sources (which are not real) based on its training, a known flaw. In terms of accuracy, GPT-3.5 often makes more basic mistakes or oversimplifications, especially in complex domains, and it might contradict itself over a long session. Therefore, for factual reliability, GPT-3.5 is the weakest of the three, and any important research with it would require heavy user fact-checking.

User Interaction Flow and Guidance

- **ChatGPT Deep Research:** Uses a **distinct interaction workflow** compared to normal chat. The user starts by selecting the Deep Research mode and entering a single, high-level prompt describing their research need. The AI may **ask clarifying questions** at the outset if the prompt is ambiguous or too broad (to refine the task). Once it understands the request, it initiates an autonomous research session. At this point, the process is largely hands-off: Deep Research works in the background for several minutes (often 5–30 minutes depending on complexity) and **provides a live progress sidebar** or log of what it's doing. This might show which subtopics it's investigating or what sources it's pulling from, giving the user transparency into the steps taken. The user does not need to prompt further; the agent decides the next steps on its own. When finished, it delivers the final structured report as a message in the chat. Notably, the output often includes a brief **summary of its reasoning or**

approach, explaining how it approached the query. This is very different from the standard Q&A flow – it's more like delegating a research task and waiting for the report. The user interaction is minimal during the run (aside from watching progress or answering initial clarification), which is convenient for complex tasks but requires patience for the result.

- **GPT-4 (standard ChatGPT):** Follows the **traditional turn-by-turn chat** interaction. The user typically asks a question or gives an instruction, and GPT-4 responds within seconds. For complex tasks, the user might have to break the job into parts and ask multiple questions sequentially, guiding the model through the research step by step (since GPT-4 won't autonomously decide to do multi-step research). GPT-4 generally does *not* ask the user follow-up questions – it tries to interpret whatever was asked and answer in one go. Only if the query is extremely unclear might it ask for clarification, but this is rare. There is **no persistent "plan" or background process**: each response is generated on the fly based on the current conversation state. The user stays in control of the flow, deciding the next query or which detail to probe. In essence, GPT-4's interaction is highly responsive and interactive, but the *user must micromanage* a complex research task, since the model won't self-direct beyond the current prompt. There's also no built-in interface to show what tools or steps GPT-4 is taking – because standard GPT-4 isn't actually executing tools or browsing. It's a straightforward question-answer dialog from the user's perspective.

- **GPT-3.5:** Very similar interaction to GPT-4 in that it's a **classic chat Q&A** model. The user asks a question, GPT-3.5 answers almost immediately. If the task is complicated, the burden is on the user to break it down – GPT-3.5 will not plan out multiple turns of reasoning by itself. In fact, GPT-3.5 is even more likely to *misinterpret* a vague query rather than ask for clarification, often giving a partial or generic answer. This means the user might have to re-ask or clarify manually. GPT-3.5 won't volunteer clarifying questions; it generally treats the last user prompt as complete. Additionally, because it lacks the depth of GPT-4, the user interaction might involve more back-and-forth to get to a detailed answer (for example, asking a follow-up because the first answer was too superficial). There's no progress indicator or summary of steps – each turn is independent, and any notion of a "research process" has to be managed by the user through successive prompts.

Comparison Summary Table

To summarize the differences, the table below highlights how ChatGPT's Deep Research capability compares to the standard GPT-4 and GPT-3.5 models across these key areas:

Featu	ChatGPT Deep Research	GPT-4 (Standard [illegible])	GPT-3.5 (Standard [illegible])
Web Acces s & Searc	Integrated autonomous web browsing and real-time search. Can crawl and read online content, including	No built-in web access (knowledge is limited to training data; requires plugins or external tools for	No internet access at all; entirely limited to pre-2021 training knowledge.
Multi-Step Resea rch	Yes – **conducts multi-step investigations** automatically. Breaks tasks into sub-tasks, performs iterative searches,	Limited – can handle complex queries in one response, but *does not autonomously iterate*. Any	No – answers one question at a time. Any complex research must be manually split into steps by the user (the model

Structured Output	**Structured reports by default**. Outputs are lengthy, organized documents with sections, headings, bullet points, and tables/graphs	Generally gives an essay-style answer or list. Can produce structured formats *if specifically instructed*, but not automatic. Default output	Tends to give brief, plain answers. Can do simple lists or a basic outline if prompted, but not capable of lengthy, well-structured reports by
Sources & Citations	**Cites sources for claims**, providing transparent references. Draws directly from those sources, which are listed or linked in the report. This increases trustworthiness and allows verification of each fact.	No automatic citation of sources. Answers are unsourced, based only on training data. The user must trust the answer or do their own fact-checking. Less prone to hallucinations than GPT-3.5, but can still	No citations; does not provide sources. Prone to **hallucinations and errors** on complex or novel questions. Often needs user fact-checking. If asked for references, it may even invent some, since it cannot actually
Reliability & Accuracy	High on well-sourced topics. Uses a advanced reasoning model to reduce errors and cross-verify info from multiple sources, leading to fewer simple mistakes. However, accuracy is bounded by what it finds online – if the web content is wrong or if it misinterprets	Very knowledgeable within its training scope. More accurate and detailed than GPT-3.5. Rarely makes grammar or logic errors, but **has training cutoff** (so anything truly new or updates post-cutoff are unknown). Can be confident but wrong in areas it lacks	Moderate reliability on common knowledge, but **noticeably less accurate** on detailed or technical questions. May falter on multi-step logic and often gives overly general answers if unsure. High chance of minor factual errors or inconsistent details, especially
User Interaction	**Agentic, one-shot request**: User gives a single prompt (after selecting Deep Research mode). The system might ask a couple of clarifying questions, then handles the rest autonomously. Progress is shown via a live step-by-step log (so user sees it searching,	**Interactive dialogue**: User and GPT-4 go back-and-forth. The user typically breaks complex tasks into smaller queries. GPT-4 responds almost instantly each time. No background processing or persistent plan – the conversation is driven turn by turn by the user's	**Interactive dialogue, but less proactive**: Works in the same Q&A style as GPT-4, answering each prompt briefly. GPT-3.5 is even more reliant on the user to steer the conversation and often won't seek clarification if a query is vague. The user must check its answers and ask further

Sources: The above comparison is based on OpenAI's official description of *Deep Research*, external evaluations and reports, and documented testing of these models. Each model's characteristics were derived from these sources and known usage of GPT-4 and GPT-3.5. The Deep Research mode clearly extends ChatGPT's capabilities in autonomy, depth, and formatting, whereas GPT-4 and GPT-3.5 require more user involvement and have more limited scope in comparison. The result is that Deep Research can provide a well-documented, thorough analysis (with evidence) on a complex query, versus the older models which provide quicker, simpler answers without direct source backing.