ChatGPT

Prompt Engineering in HEOR: Unlocking AI for Health Economics and Outcomes Research

Introduction

Prompt engineering – the art and science of designing effective inputs for generative AI – has become a critical skill for health economics and outcomes research (HEOR) professionals. Generative AI models like large language models (LLMs) are increasingly used to **sift scientific literature, draft documents, and even assist in study design** in the life sciences [1]. Mastering how to craft prompts allows HEOR experts to harness these powerful tools to **enhance productivity, accuracy, and innovation** in their work. For instance, major organizations have started investing in employee training on AI and prompt design; Johnson & Johnson recently trained over 14,000 staff in AI and prompt engineering basics [2]. This reflects a broader trend – effective prompting is now seen as essential for leveraging AI in complex, regulated fields like pharmaceuticals and health research.

At its core, prompt engineering means communicating with an AI in a *precise and strategic way* so that it produces useful, relevant outputs. In the HEOR context – encompassing tasks such as systematic literature reviews (SLRs), economic modeling, real-world evidence (RWE) generation, and health technology assessment (HTA) – well-crafted prompts can dramatically speed up workflows while maintaining quality. **HEOR analysts can use prompts to have LLMs summarize vast evidence, check modeling logic, draft reports, and more**, saving time on labor-intensive tasks. The payoff is significant: time-consuming processes (like a full systematic review, which can take many months [3]) may be accelerated by AI assistance [4], and routine writing or data extraction tasks can be automated to free up experts for higher-level analysis [5]. However, applying generative AI in these high-stakes areas also demands caution – outputs must be *accurate, unbiased, and compliant*. This introduction provides an overview of why prompt engineering matters for HEOR, followed by detailed applications, best practices, example prompts, and common pitfalls to consider.

Why Prompt Engineering Matters in HEOR

HEOR professionals deal with complex analyses and large bodies of evidence to inform healthcare decisions. Generative AI – when guided by strong prompts – can serve as a *force multiplier* in this work. By clearly instructing an LLM, an analyst can quickly obtain: concise summaries of research findings, drafts of sections of an economic model report, ideas for structuring an analysis, or even assistance in writing plain-language explanations of technical results.

This matters because **efficiency and accuracy are paramount** in HEOR. For example, conducting a systematic literature review requires screening thousands of references and extracting data methodically, which is highly time- and resource-intensive [3]. Recent studies show LLMs can support **many steps of an SLR process** – from literature search to study selection to data extraction [6] – under the guidance of appropriate prompts. In fact, one 2025 scoping review found that LLM-based approaches were tested in 10 of 13 typical review steps, with particularly promising results in automating literature search (in 41% of studies), citation screening (38%), and data extraction (30%) [6]. About half of those studies judged the LLM assistance as *promising* (improving efficiency without major

loss of quality) [7]. These early results underscore that well-directed AI can streamline evidence synthesis in HEOR, helping researchers cope with information overload.

Beyond literature reviews, prompt engineering enables **on-demand knowledge extraction and content generation**. Instead of manually poring over reports or trial data, an HEOR analyst can prompt an LLM with a question and get a synthesized answer. For instance, researchers can pose complex, open-ended questions to an LLM (e.g. asking it to summarize all known outcomes of a treatment) and *iteratively refine the prompt to integrate additional evidence*, all without needing to re-train the model [8]. LLMs' ability to understand natural language questions means that with the right prompt, they can pull out insights from unstructured text that would otherwise require extensive human effort. This capability, when applied carefully, boosts productivity – analysts spend less time on menial transcriptions or first-draft writing, and more time on interpretation and decision-making. Indeed, companies like Merck are leveraging internal GPT-based tools to **auto-generate first drafts of scientific and regulatory documents**, freeing scientists from tedious writing chores and accelerating report preparation [5].

Finally, prompt engineering is crucial for **maintaining quality and compliance** when using AI in regulated, high-stakes environments. A poorly phrased query to an AI can yield irrelevant or even incorrect information – clearly a risk when dealing with health economic evidence that might inform policy or pricing. By mastering prompt design, HEOR experts can minimize these risks: they learn to provide context and constraints in the prompt so that the AI's output stays on-topic and factual. For example, **the phrasing of a prompt can significantly affect an AI's answer**. In one study, medical researchers used ChatGPT on real hospital patient data to identify drug–drug interactions; they found that the model's accuracy varied widely depending on how the query was worded, illustrating that subtle prompt tweaks could change the outcome [9]. In fields like healthcare where precision is critical, such variations can have *far-reaching consequences* [10]. Prompt engineering, therefore, is not just a productivity hack – it's a necessary discipline to ensure **AI-driven analyses are reliable, bias-aware, and safe for use in healthcare decisions**.

Applications of Prompt Engineering in HEOR

Systematic Literature Reviews and Evidence Synthesis

Conducting systematic literature reviews (SLRs) is a foundational activity in HEOR, used to compile all relevant evidence on topics like clinical efficacy, quality of life, and cost-effectiveness. It's also an area where prompt-engineered AI assistance is rapidly emerging. **Large language models can be directed via prompts to automate or accelerate many SLR tasks**. For example, an analyst might use an LLM to screen study titles and abstracts for relevance, extract data points from articles, or even draft summaries of findings. With proper prompt design, an LLM can operate in a *zero-shot or few-shot* mode to classify study abstracts against inclusion criteria **without continuous human input** [11]. This means if you describe the inclusion criteria clearly in the prompt (and perhaps give one or two example inclusions/exclusions), the model can sift through text and label studies as relevant or not. Likewise, you can prompt the AI to **extract key outcomes or patient population details** from an abstract by explicitly requesting those details in your input.

However, to integrate an LLM into a systematic review workflow, one must plan the process carefully. The AI should be used to augment, not fully replace, human researchers – at least given the current state of the technology [12]. A recommended approach is to divide the review process into stages and determine how prompts and AI fit into each stage alongside human oversight. **Figure 1** below illustrates a simplified workflow from a recent study on using LLMs for literature screening. It breaks the

process into four phases: data preparation, model/prompt configuration, screening, and finalization (quality control).

Figure: Simplified workflow for integrating LLMs into a systematic review screening process. The process is divided into four phases – (1) data preparation (gathering references and removing duplicates), (2) model and prompt configuration (choosing an appropriate LLM and formulating clear screening instructions), (3) screening (AI performs automated abstract classification, followed by human verification of AI decisions), and (4) finalization (human reviewers conduct quality control, refine prompts if needed, and synthesize the included studies into the final review) [13]. Such frameworks show how prompt engineering and human expertise work hand-in-hand: the **prompts must be carefully crafted and tested** in phase 2, and any questionable AI outputs are caught in phase 4 before publication.

In practical terms, here are some ways HEOR teams are using prompt-driven AI in evidence synthesis:

- **Literature Search Assistance:** By providing a prompt like *"Find key themes in the literature on [intervention] for [condition]"*, an LLM can summarize trends from a large set of article titles or snippets. While the AI won't replace a formal database query, it can help prioritize areas or suggest synonyms to broaden the search.
- **Citation Screening:** An LLM can triage abstracts if prompted with clear inclusion/exclusion rules. For example, one could feed an abstract text into a prompt: *"Determine if the following study is an RCT in adults and pertains to quality-of-life outcomes for [disease]. Answer 'Include' or 'Exclude' and cite supporting text."* Early research indicates LLMs can achieve reasonable sensitivity in screening [6], though human reviewers should double-check borderline cases (and the AI should be instructed to show *why* it included or excluded, to catch any misunderstandings).
- **Data Extraction:** You might prompt the AI to pull specific results from a paper. For instance: *"From the text below, extract the sample size, intervention, comparator, primary outcome, and reported effectiveness measure."* The AI will output the requested items, which a human can verify against the source. This can save time in building evidence tables.

Example Prompt – Systematic Review Screening: *"You are a researcher screening studies for a review on Drug X for Disease Y. Read the abstract below and decide if the study meets these criteria: adult patients with Disease Y, intervention is Drug X (alone or in combination), and outcome includes quality-of-life or economic endpoints. Respond with 'Include' or 'Exclude' and a one-sentence justification using info from the abstract."*

In this example, the prompt clearly defines the task (screening for specific criteria) and even sets the role ("you are a researcher") to orient the LLM. It asks for a specific format (decision plus justification). Such clarity helps prevent the AI from wandering off-topic. Notably, **prompting the model to provide a justification with evidence can guard against errors** – if the AI must quote or reference the abstract to explain its decision, it's less likely to hallucinate study details. Research suggests that requiring models to cite supporting text is a useful strategy to minimize unwarranted assumptions [14]. In fact, one study proposed having LLMs quote verbatim sentences supporting their classification, which made it easier for humans to verify the AI's reasoning and catch hallucinations [14].

Overall, prompt engineering in SLRs is about finding the sweet spot between **automation and oversight**. When done right, it can dramatically cut down the screening and summarizing workload [15], while still ensuring that final inclusion decisions and syntheses remain high quality. Current best practice is to use the AI as an assistant – e.g. to flag likely inclusions, draft summaries, highlight data – and have the human reviewers quickly review these outputs. As evidence of the potential, a 2025 review concluded that although fully validated LLM solutions for reviews are still lacking, their rapid development and positive initial results **"highlight [LLMs'] growing relevance"** in evidence synthesis

tasks [16]. We can expect prompt strategies for literature reviews to keep improving as more researchers experiment and share what works.

Economic Modeling and Analysis

Health economists build decision models – like cost-effectiveness models, budget impact models, and epidemiological simulations – to predict outcomes and inform resource allocation. Developing these models and communicating their results is another area where prompt-engineered AI support can be valuable. While an LLM won't be running complex math (and we wouldn't trust it to do precise quantitative calculations), it **can assist with the many text-based aspects of economic modeling**.

Consider the lifecycle of a pharmacoeconomic model: defining the model structure, documenting assumptions, writing the analysis plan, coding parts of it (in pseudo-code or actual code), explaining results in plain language, and preparing reports or publications. Each of these steps has a *language component* that an LLM can help with if prompted effectively:

- **Model Conceptualization:** You can use prompts to brainstorm or refine the model structure. For example, *"Outline a Markov model for Disease Y with three states (healthy, diseased, dead) and describe the transitions and cycle length assumptions."* The AI might produce a reasonable initial outline of health states, transitions (e.g. progression, mortality), and say something about cycle length (e.g. "one month cycles, with transition probabilities estimated from literature"). The human modeler can then take this outline and adjust it with domain knowledge. The prompt essentially helps generate a draft concept which can spur further thinking.
- **Parameter Identification:** Suppose you're not sure what inputs you need for a cost-effectiveness model. A prompt like *"List the key parameters required for a cost-effectiveness analysis of Drug X vs standard of care in Disease Y (e.g. clinical, utility, cost inputs)"* will likely yield a checklist: incidence of Y, drug efficacy rates, utility values for health states, costs of drug and disease management, discount rate, time horizon, etc. This can serve as a starting template to ensure you don't overlook something important.
- **Technical Writing and Explanation:** A large part of modeling is writing – describing methods and results. You can direct the AI to *draft pieces of the report*. For instance, *"Explain in simple terms the meaning of an incremental cost-effectiveness ratio (ICER) and how it was interpreted in this study's results."* This prompt would get ChatGPT to produce a lay-friendly explanation of an ICER, which you could then fact-check and refine. It's a quick way to generate patient or stakeholder-friendly text. Similarly, after obtaining model results, one could prompt the AI: *"Summarize the model findings: Drug X had an ICER of $50,000/QALY compared to standard care; explain whether this is considered cost-effective under common willingness-to-pay thresholds."* The AI might contextualize the number (e.g. "This ICER is around the threshold of $50K per QALY often cited in literature, meaning Drug X may be on the borderline of cost-effectiveness."). Of course, the analyst must verify and correct any misstatements, but the prompt-guided output provides a helpful first draft.
- **Code Assistance:** While not a primary focus of prompt engineering as discussed in this course, it's worth noting that advanced models (especially those like GPT-4 with coding capabilities) can assist in writing code for models if prompted. For example, *"Here is a pseudo-code of a Markov model, identify any errors or suggest improvements,"* or *"Write a Python function to calculate the QALYs given a list of utility values and cycle lengths."* The AI can often generate functional code or debug issues, which the health economist can then implement. This kind of AI-assisted coding can speed up model implementation, though caution is needed to ensure the code is correct and efficient.

It's important to stress that **human expertise remains central in economic modeling**, even with AI help. The role of prompt engineering here is to handle the boilerplate and provide creative support. The economist defines the problem and verifies the answers. For example, an AI might not automatically know all the nuances of a given disease model (like why a particular half-cycle correction is needed), but if you *tell it explicitly in a prompt* ("assume half-cycle correction is applied"), it will incorporate that into its explanation. The onus is on the user to supply accurate context.

One notable benefit is improved **communication of model results**. HEOR often requires translating technical findings into insights for decision-makers (who may not be economists). With prompt engineering, you can generate multiple versions of an explanation tuned to different audiences. A prompt could be, *"Explain these results to a hospital formulary committee: [paste model outcome highlights]."* The AI might focus on budget impact and clinical implications in non-technical language. Another prompt could target a scientific audience: *"Explain the same results to an academic conference, highlighting methodology and uncertainty."* By adjusting the prompt, the model will shift the tone and detail level. This capability to **quickly repackage information** is extremely useful for market access and dissemination. It unlocks a form of agility – you have a "junior writer" on demand that can adapt your core results into various formats (each of which you will edit and fact-check, of course).

In summary, prompt engineering can support economic modeling by **accelerating documentation and clarifying complex concepts**. It helps ensure no steps are overlooked (via checklists and outlines) and that the findings can be communicated effectively to stakeholders. Just as importantly, it can serve as a real-time sounding board – if you prompt an LLM with "What are potential limitations of this model?" after describing it, the AI might list common issues (e.g. "uncertainty in long-term extrapolation" or "data is from trial not real-world") which you can then address or mention in your report. Used in this interactive way, an AI can mimic a colleague giving feedback. Such uses show the **innovative potential** of prompt engineering in HEOR: it's not merely about getting the right answer, but about exploring the problem space more thoroughly and communicating insights more clearly.

Real-World Evidence Generation and Analysis

Real-world evidence (RWE) refers to data and insights gathered from real-world settings – for example, observational studies, electronic health records, registries, claims databases, and patient-reported outcomes. Generative AI can play a role in **making sense of messy real-world data** and aiding the generation of evidence from it, with prompt engineering guiding the process.

One immediate application is in **summarizing and interpreting RWE study results**. RWE studies often yield nuanced findings (e.g. effectiveness of a drug in routine practice, or safety signals in a broader population) that must be distilled into key messages. An HEOR specialist can leverage prompt engineering to summarize lengthy real-world study reports. For example: *"Summarize the following real-world study results of Drug X in 2-3 sentences, focusing on effectiveness and any safety concerns compared to clinical trial data."* The model, given the study's results text, could output a concise summary like, "In a real-world cohort of 5,000 patients, Drug X reduced hospitalization rates by ~15% relative to standard treatment, though the absolute benefit was smaller than seen in RCTs. No new safety issues emerged, but adherence was lower in routine use, potentially impacting outcomes." Such a summary provides a quick takeaway that the analyst can then refine. By prompting the AI to compare to clinical trial data, we ensure it frames the RWE in context, which is often what stakeholders need to know.

Another use of prompting is **protocol or analysis plan drafting for RWE studies**. If an HEOR team is designing a new observational study, they might prompt the AI for components of the protocol: *"Draft an objective and study design paragraph for an observational study of Drug X in Disease Y using a claims database. Include the population, time frame, and outcomes of interest."* The AI can produce a starting

paragraph: “This study will retrospectively evaluate the effectiveness of Drug X in patients with Disease Y using the [Country] claims database, from 2018–2023. We will identify adults with at least one prescription of Drug X and a diagnosis of Disease Y, and follow them for outcomes including all-cause hospitalization, disease-specific complications, and healthcare costs. The primary objective is to compare these outcomes between Drug X users and matched patients on standard therapy, using propensity score matching to control for confounding.” A human researcher would fine-tune this, but having a draft ready in seconds is a significant head start. It ensures key elements are mentioned and can prompt the team to consider pieces they might have missed (e.g. matching method, time frame) if the AI includes them.

LLMs can also assist with **data cleaning and coding** tasks in RWE, indirectly via prompting. For instance, if given a data dictionary or some sample data, you could ask, *“How can I handle missing data in this dataset of heart failure hospitalizations? Suggest some methods.”* The AI might reply with techniques like multiple imputation, last observation carried forward (if longitudinal), or simply state that you should do a sensitivity analysis with different assumptions. While not a direct analysis, this advice (which comes from the model’s training on statistical knowledge) can be helpful to the analyst as a thought partner.

A particularly novel application is using AI to **generate synthetic patient scenarios or narratives**. For example, to humanize the data, one might prompt: *“Generate a hypothetical patient case that illustrates the benefit of Drug X in a real-world setting, based on the study findings.”* The model could produce a brief story: “*Patient A is a 60-year-old with Disease Y who started Drug X last year. In the 12 months before Drug X, they had three hospital visits. Since starting Drug X, they’ve had only one minor urgent care visit and report improved daily functioning…*” etc. This kind of narrative can be useful in presentations or reports to complement the statistics with a tangible example. Of course, it must be clearly labeled as a hypothetical scenario (not an actual patient), but it’s a way to make RWE more relatable. Prompt engineering is key to getting a relevant scenario (“based on the study findings” was included to nudge the AI to reflect the data trend of reduced hospitalizations).

It should be noted that **RWE often involves sensitive data** (patient information) and complex causality issues. Directly using an AI on raw patient-level data is generally not feasible due to privacy (and because LLMs are not designed to parse large structured datasets easily). However, summarizing aggregated results or querying the AI for medical knowledge to interpret results is within reach. For example, if an observational study finds a certain subgroup has better outcomes, one might ask the AI: *“What are some possible reasons why younger patients respond better to Drug X than older patients, as observed in this study?”* The AI might bring up metabolism differences, fewer comorbidities, better adherence, etc., which can help the researchers ensure they’ve considered those angles in their discussion.

Finally, generative AI can support **market access teams using RWE** by helping craft the narrative for payers. Market access often uses RWE to demonstrate real-world value of an intervention. Through prompt engineering, you can tailor messages: *“Explain to a healthcare payer how the real-world data on Drug X supports its cost-effectiveness – focus on reduced hospitalization and total cost findings.”* The AI’s response might say, “Real-world data show that Drug X is associated with a 20% reduction in hospitalizations among patients with Disease Y. For payers, this means substantial cost savings: approximately \$2,000 less in annual healthcare costs per patient compared to those not on Drug X [1]. These savings offset the higher drug price, leading to an overall cost-effective profile in routine care.” The reference to costs and savings is exactly the kind of framing payers expect. By refining the prompt, you ensure the AI hits those points (e.g. instruct it to mention cost savings explicitly).

In sum, prompt engineering in the RWE domain helps **transform data into insights and compelling narratives**. It augments the researcher's ability to interpret and communicate real-world data by providing quick drafts, explanations, and even creative ways to present evidence. As with other applications, the outputs must be validated (e.g. ensure the numbers cited by the AI match the actual study results – an AI might sometimes mis-remember or hallucinate figures if not directly provided). One study highlighted that generative models may confidently produce unsupported statements if asked to "infer" data not given [17] – a cautionary tale for RWE where the model should stick to observed results. The solution is to craft prompts that keep the AI grounded (e.g. by prefacing with known facts or explicitly telling it not to make up data). With careful prompting and oversight, LLMs can become powerful assistants in turning real-world data into real-world evidence.

Health Technology Assessment and Market Access

Health Technology Assessment (HTA) is a formal process of evaluating the clinical, economic, and patient-impact evidence for a health technology (like a new drug, device, or intervention) to inform policy and reimbursement decisions. Market access professionals prepare extensive dossiers and submissions for HTA agencies and payers, which often run hundreds of pages covering clinical study results, economic models, budget impact analyses, and more. **Prompt engineering offers valuable support in creating and refining this content**, as well as in generating examples and arguments to strengthen market access communications.

A key challenge in HTA submissions is articulating a clear and compelling value story for the new technology. This involves weaving together data from clinical trials, real-world studies, and economic analyses into a coherent narrative. Generative AI can be prompted to **draft sections of an HTA dossier** or to **simulate critical review questions** that an HTA committee might ask. For instance:

- **Drafting Clinical and Economic Summaries:** You could prompt an LLM to write a summary of the clinical efficacy data for the submission. *"Summarize the clinical effectiveness of Drug X for Disease Y based on phase III trial results, highlighting the key outcomes (e.g. survival, symptom improvement) and their statistical significance."* The model will produce a paragraph that you can then fact-check against the trial data. Similarly, an economic summary can be drafted: *"Summarize the cost-effectiveness results of Drug X vs standard care from the model – include the base-case ICER and mention any sensitivity analysis findings that are important."* This yields a starting summary like, "Base-case analysis shows an ICER of \$45,000 per QALY gained for Drug X versus standard care [18]. This result is robust in sensitivity analyses, with no scenario exceeding \$60,000/QALY, indicating a high likelihood that Drug X is cost-effective under common willingness-to-pay thresholds." If the AI doesn't automatically include certain points (e.g. budget impact or key assumptions), you can refine the prompt or simply add those details manually. The benefit is that the bulk of the phrasing is done, and it's easier to edit than to write from scratch. Notably, some pharma companies have begun using GPT-based tools to **generate first drafts of regulatory and HTA documents**, precisely to save time on writing and allow experts to focus on review [5] [18].

- **HTA Query Simulation:** Prompt engineering can help anticipate tough questions. An example prompt might be: *"You are an HTA reviewer evaluating Drug X. List three critical questions or concerns you might raise after reviewing the submitted evidence (clinical and economic)."* The AI might output, for example: "1) Is the comparative efficacy of Drug X vs its main competitor clinically meaningful or just statistically significant? 2) What is the uncertainty in the long-term cost-effectiveness given limited trial follow-up? 3) How do we know the real-world adherence to Drug X won't be lower, eroding its benefit?" These are realistic questions. By getting these from the AI, the market access team can ensure their submission or responses address them. Essentially, the AI, guided

by the prompt, role-plays as a skeptical HTA committee member. This use of prompts can significantly improve preparedness for actual HTA meetings or written responses.

- **Creating Lay Summaries and Value Propositions:** Many HTA agencies require a plain-language summary for the public or have sections where the value proposition needs to be stated succinctly. With prompt engineering, you can instantly toggle the complexity of language. For example: *"In one sentence, explain why Drug X should be reimbursed, focusing on what makes it valuable to patients and the health system."* An LLM might respond, "Drug X offers a significant improvement in patient survival and quality of life at a cost that is justified by these health gains, meaning it provides good value for the money in treating Disease Y." This one sentence, while simple, captures a value proposition. You might then expand it, or use it as the final line in an executive summary. Because LLMs are good at summarization, they often excel at these "bottom line" statements if the prompt clearly asks for them.

- **Adapting Content for Different Stakeholders:** Market access often involves tailoring the message to different audiences – e.g., a slide deck for clinicians, a briefing for payers, a press release for general public. Starting from the same evidence base, you can use prompts to shift emphasis. For clinicians: *"Summarize the benefits of Drug X for a clinical audience, focusing on patient outcomes and safety."* For payers: *"Summarize the benefits of Drug X for a payer audience, focusing on cost savings (like reduced hospitalizations) and cost-effectiveness."* For patients: *"Summarize in plain language how Drug X could improve a patient's life."* In each case, the LLM will likely pick different facets to highlight (efficacy vs economic vs quality of life) based on the prompt cues. This rapid re-targeting of content is incredibly useful in practice, as it ensures consistency across materials while speaking the language of each stakeholder.

Example Prompt – HTA Value Summary: *"You are a health technology assessment analyst preparing a recommendation. Summarize the value of Drug X for Disease Y in 5 bullet points, covering: (1) Clinical efficacy advantages, (2) Safety/tolerability, (3) Impact on quality of life, (4) Cost-effectiveness, and (5) Budget impact for the healthcare system."*

In this prompt, by specifying the structure (5 bullet points) and the topics for each bullet, we guide the AI to produce a comprehensive value summary. The output might be: (1) Drug X significantly improves progression-free survival by 6 months compared to current therapy [18]. (2) It has a comparable safety profile with no increase in serious adverse events. (3) Patients on Drug X report better quality of life, with improvements in fatigue and daily functioning. (4) Drug X is cost-effective, with an ICER of \$45K/QALY, under common willingness-to-pay thresholds. (5) The budget impact is moderate – estimated at 0.5% of the total drug budget in year one, offset by savings from fewer hospitalizations. – This kind of output hits all the key points in a very digestible format. Even if the AI doesn't get every detail perfect (we'd verify each point against the actual analysis), it provides a strong starting draft to refine. It's easy to see how such prompt-engineered content can accelerate the assembly of an HTA dossier or a payer submission, which normally requires synthesizing information from many sources.

That said, **caution is critical**. All content produced must be verified against source data and vetted for compliance. HTA submissions undergo scrutiny, and any factual error or unfounded claim can undermine credibility. Common pitfalls like hallucination or bias in AI output could be especially problematic here – for example, if the AI mistakenly cites a study result that doesn't exist or phrases something in a way that overstates the benefit. Thus, prompt engineering for HTA is often done with a conservative approach: you provide the model with the data (embedding actual numbers or findings in the prompt whenever possible) and ask it to rephrase or organize that data, rather than asking it to pull from memory or guess. Additionally, companies are adopting secure, internal LLM solutions for such uses due to confidentiality. One survey noted that many pharma companies banned employees from

using public ChatGPT for fear of leaking confidential pipeline or pricing information [19]. In response, firms like Merck and Lilly have built internal GPT platforms so that sensitive prompt content (like details of a not-yet-approved drug's dossier) stays in-house while still enabling the efficiency gains [5] [18].

In conclusion, prompt engineering can significantly enhance the **speed and clarity of HTA and market access work**. It helps writers generate high-quality drafts, anticipate questions, and tailor communications – all crucial in getting a new health technology successfully through the reimbursement and uptake process. By embracing these AI tools with proper guidance and oversight, HEOR and market access teams can improve both their productivity and the quality of their deliverables.

Best Practices for Effective Prompt Engineering in HEOR

Applying prompt engineering in HEOR tasks requires not only creativity but also rigor. Below are several best practices and principles to ensure you get the most out of LLMs while maintaining accuracy and compliance:

- **Be Clear, Specific, and Directive:** The prompt should precisely state what you want. Ambiguous prompts yield ambiguous answers. Include relevant details such as the context (e.g. "you are an HEOR analyst" or "for a payer audience"), the desired output format (bullet points, summary paragraph, etc.), and any constraints (e.g. "use layman's terms" or "cite supporting evidence"). For example, instead of asking *"What is the outcome of the study?"*, ask *"Summarize the primary outcome of the study X in one sentence, including the magnitude of the effect."* This removes guesswork for the AI and improves the relevance of its response.

- **Provide Context or Data When Possible:** LLMs work best when they have the necessary information. In HEOR, that means feeding the AI the snippets of source material you want it to use, rather than expecting it to recall specific details from memory (which may be outdated or incorrect). For instance, if you want a summary of a trial, give the key results in the prompt and ask for a summary of those. Example: *"Drug X reduced A1c by 1.2% (95% CI 0.8–1.6) vs placebo in a 52-week trial. Now summarize this result in one sentence highlighting its significance for patients."* By providing the data, you reduce the chance of the model hallucinating numbers or facts. This approach aligns with the idea of **retrieval-augmented generation**, where the AI is grounded in supplied evidence [20]. In practice, always prefer to supply a chunk of a paper or a data table into the prompt, then ask questions about it, rather than asking the AI to recall facts on its own.

- **Use Structured Prompts for Complex Tasks:** If the task is multi-step or complicated (like generating an HTA summary covering multiple domains), consider breaking the prompt into steps or using numbering as we did in the HTA bullet example. You can also prompt iteratively: start with one prompt to get a baseline, then refine. For example, *first prompt:* "Draft a paragraph on the cost-effectiveness of Drug X versus Y." *Follow-up prompt:* "Now add a sentence about uncertainty and one about budget impact to that paragraph." Chaining prompts like this (sometimes called **prompt chaining**) leverages the AI's previous output and steers it gradually to the desired final output. In scenarios requiring reasoning, you might invoke a chain-of-thought approach, e.g., *"Think step by step: first list the differences between trial population and real-world population, then explain how each difference could affect outcomes."* Explicitly guiding the model's reasoning can lead to more thorough and logical responses.

- **Leverage Role-Playing:** As seen in examples, setting a role in your prompt can anchor the model's style and perspective. If you say "You are a health economist explaining to a policy maker…", the model will try to adopt that mindset, which influences the tone and content. This is

a powerful way to get outputs that are appropriate for different stakeholders. Roles like "expert clinician", "patient advocate", "statistician checking the model" can be experimented with. It's part of prompt engineering to find which roles yield the best insights for the task at hand.

- **Instruct for Format and Length:** Tell the AI exactly how you want the answer structured. If you need a numbered list, say so. If you want a short answer or a long report, include that (though models sometimes ignore length instructions if they conflict with content, they usually try). For example: "Provide three bullet points on X" or "In no more than 100 words, define Y." In HEOR writing, you might have specific format needs (e.g. PICO format for a summary of evidence – Population, Intervention, Comparator, Outcome). You can prompt: *"Summarize the trial results in a PICO format."* If the model understands, it might output something like: "**Population:** ..., **Intervention:** ..., etc." – if not, you may need to clarify in the prompt. Being explicit with format reduces editing later.

- **Iterate and Refine**: Treat the AI's first response as a draft. It's often useful to review the output and then **prompt again with adjustments**. You can feed back to the model what to fix: *"The above summary is missing mention of adverse events – please add a sentence on safety outcomes."* The model will then (usually) comply and produce a revised version. This iterative loop is where a lot of the power lies – you as the expert see what's wrong or incomplete, and use a prompt to correct it. Iteration also helps in overcoming limitations: if an output is too generic, add more detail to the prompt and run it again; if it's too verbose, instruct "make it more concise".

- **Verify Factual Accuracy and Sources:** Always double-check any factual statements the AI makes. Even with good prompts, an LLM might **hallucinate** – i.e., produce a confident-sounding claim that isn't true or wasn't in the provided data [21]. This is especially true if your prompt inadvertently encourages speculation (e.g. "infer the result" or "imagine if..."). As a best practice, avoid prompts that ask the model to guess unknowns. If you need an assumption, specify it yourself rather than leaving it to the AI. Additionally, a useful prompt strategy is to ask the model to provide sources or quotes for its statements (as we did in screening justification). For example, *"Provide the summary and list the source of each key fact (e.g. trial name or figure number)."* This can at least make the model reveal what it *thinks* is the source, which you can then verify. In one approach, researchers had the model cite verbatim text from references to support its outputs, significantly reducing hallucinations and making human verification easier [14].

- **Maintain Ethical and Compliant Use:** In regulated fields, ensure your prompt and outputs adhere to privacy and compliance requirements. **Never input confidential patient data or sensitive company data into a public AI service** – if using public models like ChatGPT, de-identify or abstract any sensitive info (or better yet, use a secured internal model). When generating content, remember that even if the AI writes it, you are responsible for it. So you need to ensure it doesn't inadvertently plagiarize or introduce bias. Keep an eye out for biased language or assumptions in outputs; models can reflect biases present in their training data [22], such as gender or racial biases, which is not acceptable in professional communications. If you spot any such bias, correct it and consider adjusting the prompt to avoid it (for instance, instruct the model to use neutral language or focus on data).

- **Use Domain-Specific Models or Tools if Available:** While general models like GPT-4 are quite capable, there are emerging tools fine-tuned for medical or scientific use (e.g. models that are less likely to produce harmful medical advice). If your organization provides a special interface (some companies integrate literature databases with GPT-style Q&A, for example), use those as they might have guardrails or up-to-date literature access built-in. When dealing with numerical data or calculations (like in economic models), consider using the AI in combination with

traditional tools: e.g., use Python/R for actual number crunching and then use the LLM to help interpret the results. Prompt engineering can thus be part of a larger workflow where AI and conventional analysis complement each other.

By following these best practices, HEOR professionals can greatly improve the quality of AI-generated outputs and reduce the time spent fixing errors. In essence, **a good prompt is like a good survey question or a well-defined analysis plan** – it yields meaningful, targeted answers. As you gain experience, you'll develop an intuition for phrasing that works best. Keep notes of successful prompt formats and reuse them. For example, you might find that saying *"list 3 strengths and 3 weaknesses of X"* always gives a balanced analysis – that can be a template. Prompt engineering is an evolving skill, so continuous learning and adaptation are part of the process. The effort you put into crafting thoughtful prompts will pay off in more efficient and insightful results from the AI.

Common Pitfalls and How to Avoid Them

While the potential of prompt engineering in HEOR is exciting, it's equally important to be aware of the pitfalls. Misuse or blind trust in AI outputs can lead to serious errors. Below we highlight common issues and how to mitigate them:

- **Hallucinations (Fabricated Information):** As mentioned, LLMs can produce text that sounds valid but is not backed by the input or reality. In a literature review context, a hallucinating AI might **confidently state a result or methodology that doesn't exist in the actual study** [21]. For example, it might wrongly assert "Study A was a randomized trial" when it was observational, or completely make up a statistic. This is dangerous in HEOR where accuracy is paramount. To avoid this: always cross-check critical facts with original sources. Structure prompts to *not encourage guessing* – e.g., ask the model to summarize given data, not to infer missing data. If you sense an output might include hallucination, you can prompt the model to double-check ("Are you sure these facts are in the text?"). Some studies suggest requiring the model to provide supporting text can reduce hallucination [17]. Ultimately, **human validation is the fail-safe** – never copy-paste an AI-generated evidence summary into a submission without verifying every datum.

- **Bias and Ethical Concerns:** LLMs learn from vast text data, and unfortunately this can include biases. They might output stereotypical or biased assumptions about populations if not careful [22]. In health economics, this could manifest subtly – e.g., assuming a certain patient group is less adherent without evidence, or using language that is not culturally sensitive. Be vigilant for such bias in AI outputs. You should actively prompt the AI to be neutral: for instance, "Provide an unbiased comparison of treatments" or "Avoid any stereotypes in the explanation." Moreover, any equity considerations (like how an intervention affects subpopulations) should be introduced by the human expert; don't expect the AI to volunteer these unless it was in the training data. Bias mitigation also involves diversity in prompts – if appropriate, test the prompt with scenarios involving different populations to see if the AI's tone or recommendations change without justification, which could signal bias. Ensuring **ongoing validation and human oversight** is key to catching biases [20].

- **Over-reliance on AI (Automation Bias):** There is a risk that users trust the AI output too much, especially if it's well-worded. In high-stakes fields like pharma, one must remember that the **LLM is not an oracle**. It doesn't actually understand the science; it's predicting likely text. Automation bias could lead an analyst to accept an AI-generated conclusion without sufficient skepticism. To combat this, treat AI as an assistant that *always needs review*. Encourage a culture of verification:

if the AI drafts an HTA section, perhaps a colleague double-checks the content independently. Use the AI to assist thinking, but critical thinking remains with the human. If an AI suggests a certain result is "significant" or a drug is "cost-saving," ensure those claims are validated by your statistical analysis or budget model. Essentially, **never let the AI have the final sign-off** on something that you wouldn't otherwise sign off on yourself.

- **Data Privacy and Confidentiality:** Prompting an AI often involves inputting text that could be proprietary or sensitive (like unpublished results, patient descriptions, etc.). A huge pitfall is inadvertently leaking this information into a system that might not be secure. Public AI platforms may store your conversation (even if not intended for malicious use, it's out of your control once on their servers). As best practice, **do not use sensitive data in prompts on external AI platforms**. If you must, anonymize and abstract it. Many organizations are addressing this by using private instances of LLMs or tools where data won't be retained. If you're unsure, err on the side of caution. Also, be mindful of intellectual property – e.g., if you ask the AI to rewrite a paragraph from a published article, ensure you're not inadvertently plagiarizing the original (the AI might output text close to the source). Always cite original sources for information, even if the AI helped summarize them, to give proper credit and uphold academic integrity.

- **Regulatory and Compliance Issues:** In pharma and health research, communications are often regulated (think of promotion rules, or requirements for balanced evidence). An AI doesn't inherently know these rules. It might draft a paragraph that accidentally makes a promotional claim not supported by data, or uses language not aligned with regulatory guidelines. For example, it might use superlatives ("Drug X is a groundbreaking cure…") which would be flagged in an official context. It's the user's responsibility to enforce compliance. One way is to include in the prompt any relevant cautions: *"Draft the text in a scientifically neutral tone without making unsubstantiated claims."* That can reduce hyperbole. Additionally, when using AI for pharmacovigilance or safety-related content, ensure it includes the necessary details (the AI might omit a black box warning unless prompted). Essentially, integrate your domain's compliance checklist into the prompt or the review process. Remember, as of now regulators do not accept "the AI wrote it" as an excuse for any oversight – the onus is fully on the submitting professionals.

- **Context Limitations:** LLMs have context length limits (they can only consider a certain amount of text in one go). If your prompt or needed reference materials are very long (e.g., trying to feed an entire 200-page report), the model may not reliably integrate everything – it might focus on the beginning and end, for instance. A pitfall is assuming the AI "read" all the long input you gave it. To avoid this, break tasks into chunks. Don't ask the model to summarize a massive document in one shot; give it section by section. If you notice an output missing something that was in the input, it could be due to context cutoff. Solve it by explicitly including that piece in a new prompt iteration. Also be mindful that if you carry a very long conversation with an AI, earlier facts can get lost (depending on how the system manages the conversation window). It can help to re-provide key info in your prompt if it's critical.

- **Illusion of Objectivity:** Because AI outputs are machine-generated, people might assume they are unbiased or correct. There's an *illusion of objectivity* – "the computer said so, so it must be true." This is dangerous if it leads to uncritical acceptance of results. Always question the AI's conclusions as you would a human's. If the AI says "Drug X is cost-effective in all scenarios," ask yourself – did it consider scenario A, B, C? If not, you should. Think of AI as a colleague who sometimes has brilliant insights and sometimes talks nonsense – you have to tell which is which. One strategy is to **cross-validate** important outputs: ask the same prompt in a slightly different

way, or ask another AI model, and see if results converge. Discrepancies might highlight areas to investigate further.

- **Unanticipated Tone or Wording:** Sometimes the AI might produce text that is stylistically off – too casual, too formal, or just awkward in phrasing. This is minor compared to factual issues, but still a pitfall for quality. It happens if the model picks up an unusual style from the prompt or the data it was given. The solution is simple: if the tone is wrong, explicitly instruct the desired tone and regenerate. For instance, "Rewrite the above in a formal tone suitable for a journal" or "Make the tone more accessible for a patient reader." Prompt engineering is as much about tone and style as content, especially in communications tasks.

To summarize, **the safe and effective use of AI in HEOR requires a mix of technical savvy and domain vigilance**. Prompt engineering can mitigate some pitfalls – e.g., by asking for supporting evidence to catch hallucinations, or by clarifying context to reduce bias – but it cannot eliminate all errors. The human expert remains the final gatekeeper. By staying aware of these common issues and putting in place checks (like verification steps, peer review of AI-assisted content, and secure data handling procedures), one can enjoy the productivity and creative gains of generative AI without falling victim to its drawbacks. In many ways, this is analogous to any new powerful tool: it expands what's possible, but you must learn to use it responsibly.

Conclusion

Prompt engineering is rapidly becoming an indispensable skill in the toolkit of HEOR professionals. As we have explored, it enables practitioners to unlock the full potential of generative AI models like ChatGPT, applying them to a range of activities – from accelerating systematic literature reviews to crafting clearer economic model communications, from interpreting real-world evidence to preparing robust HTA submissions. By harnessing the art of well-crafted prompts, HEOR teams can achieve **greater productivity** (through faster drafting and analysis), **greater insights** (through AI-augmented idea generation and summarization), and potentially **greater impact** (through more polished and audience-tailored deliverables).

Crucially, this power comes with the responsibility to uphold the **rigor and ethics** that define health outcomes research. The best outcomes arise when human expertise and AI capabilities are combined thoughtfully. An HEOR analyst provides domain knowledge, critical thinking, and oversight; the AI provides speed, breadth of knowledge, and language generation. With mastery of prompt engineering, the analyst can precisely direct the AI – much like a conductor with an orchestra – to produce the desired output while avoiding dissonance. The result is a harmonious collaboration where mundane tasks are minimized and human intellect focuses on interpretation, decision-making, and innovation.

As of 2025, we are still in the early days of applying generative AI in health economics and outcomes research. Best practices will continue to evolve. It's advisable for professionals to stay updated on methodological research (for example, new studies are constantly evaluating how well LLMs perform in evidence synthesis or what pitfalls they encounter [16] [20]). Moreover, engaging in community discussions – through ISPOR, professional networks, or workshops – can provide practical insights and use cases. Many HEOR groups are beginning to share their success stories (and failures) in using AI, which can inform your own practice.

In the near future, we can anticipate even more powerful models, possibly fine-tuned for medical or HEOR content, and better integration of AI into our everyday tools (imagine your reference manager or statistical software having a built-in AI assistant). But no matter how the technology evolves, the

principle of prompt engineering will remain central: **clear, contextual, and controlled communication** with AI to achieve a specific goal. By learning to "speak AI" effectively, HEOR professionals ensure that they – not the technology – remain in the driver's seat, applying these advanced tools in a way that is *responsible, transparent, and aligned with scientific integrity*.

In conclusion, prompt engineering tailored to the HEOR context is a game-changer that, when used wisely, can elevate the quality and efficiency of our research and analyses. It unlocks a new level of capability – allowing us to focus on what truly matters: generating evidence and insights that improve healthcare decisions and outcomes. With solid prompt strategies, awareness of pitfalls, and a commitment to continuous learning, HEOR practitioners can confidently integrate generative AI into their work, driving innovation while upholding the high standards of our field.

References: The information and examples above draw on current findings and expert commentary on the use of LLMs in research. Key sources include a 2025 scoping review in *Journal of Clinical Epidemiology* exploring LLM applications in systematic reviews (indicating LLMs can aid many review steps but need further validation) [6] [16], a 2025 *Information* journal article discussing prompt strategies and oversight for using GPT-4 in literature screening [11] [12], and industry observations on how pharmaceutical companies are adopting generative AI for content generation in compliance-sensitive contexts [5] [18]. The discussion on pitfalls is informed by known issues of AI like hallucinations [17] and bias [22], as well as recommendations for mitigations such as requiring source citation [14]. These references underscore the importance of careful, informed prompt engineering to reap benefits while managing risks in HEOR applications of AI.

---

[1] [2] [5] [18] [19] ChatGPT Adoption in the Life Sciences Industry | IntuitionLabs
https://intuitionlabs.ai/articles/chatgpt-adoption-life-sciences-industry

[3] [4] [6] [7] [16] Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review - PubMed
https://pubmed.ncbi.nlm.nih.gov/40021099/

[8] Generative AI in the pharmaceutical industry | McKinsey
https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality

[9] [10] Evaluating the capability of ChatGPT in predicting drug–drug interactions: Real-world evidence using hospitalized patient data - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11602951/

[11] [12] [13] [14] [15] [17] [20] [21] Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations
https://www.mdpi.com/2078-2489/16/5/378

[22] Systematic Literature Review on Generative AI: Ethical Challenges and Opportunities
https://thesai.org/Downloads/Volume16No5/Paper_30-Systematic_Literature_Review_on_Generative_AI.pdf