

# LLMs for Clinical Evidence: Automating Economic Dossiers

By Adrien Laurent, CEO at IntuitionLabs • 1/6/2026 • 35 min read

large language models

health technology assessment

economic dossiers

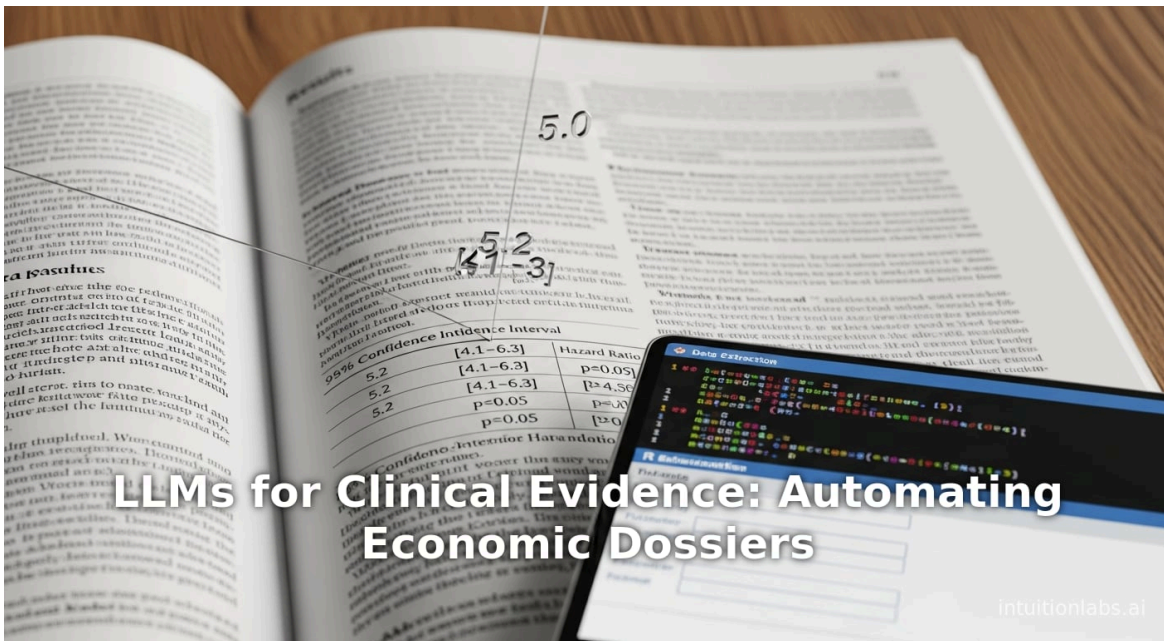
clinical evidence synthesis

systematic literature review

ai in pharma

gpt-4

ai



## LLMs for Clinical Evidence: Automating Economic Dossiers

# Using Large Language Models to Automate Collection of Clinical Evidence for Economic Dossiers and Value Stories

## Executive Summary

The process of compiling clinical evidence for economic dossiers and “value stories” is traditionally labor-intensive, requiring systematic literature reviews, data extraction, and narrative synthesis by expert teams. Recent advances in **large language models (LLMs)**, such as OpenAI’s GPT series (now at GPT-5.2), have shown substantial promise in automating parts of this process. LLMs can generate human-like text and perform targeted information retrieval tasks, enabling accelerated literature search, screening, extraction of numerical results, and even automated coding for health economic models. Pioneering studies have demonstrated LLM-driven pipelines that far outperform manual methods in speed and approach human-like accuracy. For example, TrialMind (2025) – an AI pipeline built on GPT-4 – achieved literature search recalls of 71–83% (vs. only 14–23% by humans) and improved screening rates by 1.5–2.6×<sup>(1)</sup> ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). ChatGPT-based tools have screened thousands of abstracts in hours with >95% sensitivity<sup>(2)</sup> ([bmcmmedresmethodol.biomedcentral.com](https://bmcmmedresmethodol.biomedcentral.com)) and extracted structured data from scientific tables with 100% accuracy on clearly presented values<sup>(3)</sup> ([journals.plos.org](https://journals.plos.org)), all while running 5–7 times faster than humans<sup>(4)</sup> ([journals.plos.org](https://journals.plos.org)). In health economic modeling, GPT-4 replicated published forecasts with near-perfect fidelity<sup>(5)</sup> ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov))<sup>(6)</sup> ([link.springer.com](https://link.springer.com)), suggesting that LLMs can also handle the quantitative components of dossiers.

These advances are transforming pharmaceutical value toolkit. By automating literature reviews, registries scanning, meta-analyses, and even draft narrative writing, LLMs can significantly reduce time and cost, letting experts focus on higher-level interpretation. However, major challenges remain. LLMs are prone to factual errors or “hallucinations” when data is incomplete<sup>(7)</sup> ([journals.plos.org](https://journals.plos.org))<sup>(8)</sup> ([systematicreviewsjournal.biomedcentral.com](https://systematicreviewsjournal.biomedcentral.com)). They require careful prompting and oversight, and current HTA/ **regulatory guidance** demands transparency and validation when AI is used ([www.nice.org.uk](https://www.nice.org.uk))<sup>(9)</sup> ([www.fda.gov](https://www.fda.gov)). Data privacy, model bias, and reproducibility are active concerns. Leading agencies (e.g. NICE, FDA, EMA) are issuing guidance on AI use in evidence generation, highlighting the need for audit trails, human review, and disclosure of AI methods ([www.nice.org.uk](https://www.nice.org.uk))<sup>(9)</sup> ([www.fda.gov](https://www.fda.gov)).

This report provides a detailed review of LLM capabilities, case studies, and the impact on health economics evidence generation. It covers historical context of dossier preparation, LLM methodology, specific applications (PICO search, trial screening, data extraction, economic modeling), and major research findings. Two summary tables compare performance of LLM approaches vs. traditional methods and survey published case studies. The discussion addresses regulatory perspectives, ethical issues, and future directions, concluding that LLMs hold great promise to **accelerate value dossier preparation** – as long as outputs are carefully validated by expert teams (<sup>(10)</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (<sup>(9)</sup> [www.fda.gov](https://www.fda.gov)).

## Introduction

**Health technology assessment (HTA)** and payer submissions for new therapies depend on rigorous evidence of clinical benefit and cost-effectiveness<sup>(11)</sup> ([link.springer.com](https://link.springer.com))<sup>(12)</sup> ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Pharmaceutical companies compile *economic dossiers* or *value dossiers* that include systematic reviews of trials, **real-world outcomes data**, economic models, and a narrated **value story** explaining a product’s benefits in patient and economic terms. Traditionally, experienced health economists and analysts **manually curate this evidence**: they design Boolean search strings, screen citations, extract trial data, conduct meta-analyses or population models, and write up the results for agencies. This

process is **time-consuming and resource-intensive**, often taking many months or more than a year for an entire dossier (<sup>[13]</sup> [link.springer.com](https://link.springer.com)) (<sup>[14]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). For example, **TrialMind et al.** note that systematic reviews typically require an average of 67.3 weeks and five expert reviewers to complete (<sup>[14]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Given rapid publication rates (PubMed adds ~1 million citations per year (<sup>[15]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov))), keeping evidence up-to-date strains existing methods.

**Large language models (LLMs)** such as GPT-5.2 (OpenAI), Claude Opus 4.6 (Anthropic), and Google's Gemini 3.1 Pro have emerged as a new approach. These AI systems are trained on massive text corpora and can perform complex language tasks with minimal prompting. They excel at *instruction-following*: given descriptions of tasks or data in natural language, LLMs can generate search queries, summarize information, and even write computer code (<sup>[16]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (<sup>[17]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). In HTA contexts, they have been proposed to streamline each step of evidence synthesis. For literature reviews, LLMs can **suggest search terms** from PICO elements and quickly find relevant studies; for screening, they can flag likely-included papers; for extraction, they can parse numeric results from reports; for analysis, they can even generate statistical code or run simulations (<sup>[18]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (<sup>[6]</sup> [link.springer.com](https://link.springer.com)).

Early research confirms these capabilities. A 2025 Nature **NPJ Digital Medicine** paper introduced "TrialMind", an LLM-driven pipeline for evidence synthesis. TrialMind leveraged GPT-4 prompts to generate Boolean queries, rank citation relevance, and extract trial results, integrated with human review at each step. It achieved **remarkable performance**: in a benchmark of 100 published reviews (2,220 studies), TrialMind attained literature search recalls of 0.711–0.834 versus only 0.138–0.232 for manually designed queries (<sup>[1]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (72–83% vs 14–23%). In screening, it ranked relevant studies 1.5–2.6× better than prior automated methods (<sup>[1]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). In extraction, it outperformed baseline GPT-4 by 16–32% accuracy. Ultimately, AI-assisted experts using TrialMind improved recall by 71.4% and cut screening time by 44.2%; data extraction accuracy rose 23.5% with a 63.4% time reduction (<sup>[19]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Medical experts reviewing outputs preferred TrialMind's summaries over raw GPT-4 outputs in 62–100% of cases (<sup>[19]</sup> [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). These results suggest *substantial efficiency gains* from human–AI collaboration in assembling clinical evidence.

Another example is a 2025 **PLOS One** study by Mitchell et al., which built an AI pipeline (AI-LES) using ChatGPT (gpt-3.5) to extract data for living systematic reviews. The script asked ChatGPT targeted questions about each article (e.g. "What is the estimate for the mean incubation period?"). On a test set of 94 COVID-19 papers, AI-LES extracted clearly presented study results with **100% accuracy** (<sup>[3]</sup> [journals.plos.org](https://journals.plos.org)), distinguishing means vs medians, confidence intervals, etc. It processed all 94 in 76 minutes (≈48 sec/article), whereas manual extraction took ~7× longer (<sup>[4]</sup> [journals.plos.org](https://journals.plos.org)). In aggregate, these cutting-edge studies show LLMs can handle large-scale literature tasks with accuracy comparable to humans, but in a fraction of the time.

This report comprehensively reviews how LLMs are being applied to automate clinical evidence collection for economic dossiers and value narratives. We will cover the historical need for such automation, LLM technologies, their use in specific evidence-generation tasks (search, screening, extraction, modeling, writing), and concrete case studies from literature. We also discuss regulatory/ethical considerations and future prospects. The goal is to provide an in-depth, research-based analysis of both the **opportunities** and the **challenges** of deploying LLMs in healthcare value evidence pipelines.

## Context and Background

### The Role of Clinical Evidence in Economic Dossiers

Regulators and payers (HTA bodies, insurance, government) require robust evidence when evaluating new medical products. An *economic dossier* (also called a value dossier or submission) typically includes:

- A systematic collection of clinical trial data (safety, efficacy, comparative effectiveness) from RCTs and meta-analyses.
- Real-world evidence (RWE) on usage, outcomes, and epidemiology.
- Health economic models (e.g. cost-effectiveness or budget impact models), fed by clinical input.
- A narrative *value story*: a coherent argument that integrates clinical outcomes and economic impact to demonstrate the value proposition of the therapy.

In many countries, formal HTA processes mandate submission of these dossiers. For example, NICE (UK) and PBAC (Australia) heavily scrutinize systematic review methods and modeling assumptions. Even in markets with less formal HTA, companies prepare dossiers proactively. As one analyst noted, evidence-based evaluation is “widely understood” worldwide, and even in places without formal ICER rules (e.g. Russia, parts of Asia), *economic dossiers* are “important bargaining chips” in pricing discussions (<sup>[20]</sup> [pharmafile.com](#)).

Preparing a high-quality dossier is thus critical for market access, pricing, and reimbursement. It involves **comprehensive evidence synthesis**. Analysts craft detailed PICO (Population, Intervention, Comparator, Outcome) searches, screen thousands of abstracts, extract study parameters (e.g. relative risk, survival curves), and summarize findings. Each element of the value story (e.g. “this drug improves overall survival by XX months with cost offset Y”) must be backed by data. The evidence is often heterogeneous (multiple trials, endpoints, subgroups), requiring meta-analytic techniques. Budding methodologies (e.g. network meta-analyses) are used when head-to-head trial data are lacking, combining direct and indirect evidence.

This process is slow and expert-driven: for example, Booth et al. note that typical HTA submissions take *over a year*, with systematic review and meta-analysis phases taking several months each (<sup>[13]</sup> [link.springer.com](#)) (<sup>[21]</sup> [link.springer.com](#)). Three independent analysts are often involved just to double-check screening and extraction (<sup>[22]</sup> [link.springer.com](#)). The result is that even after drug approval, access delays occur as payers vet the evidence. There is growing recognition that faster, more efficient methods are needed: a quicker evidence pipeline could mean **faster patient access** to innovations (<sup>[21]</sup> [link.springer.com](#)).

## Advent of Automation and AI

Over the past decade, major strides have been made toward automating parts of systematic reviews. Techniques like machine learning for citation screening (e.g. Abstrackr, Covidence) can prioritize abstracts (<sup>[23]</sup> [journals.plos.org](#)) (<sup>[24]</sup> [systematicreviewsjournal.biomedcentral.com](#)). Information extraction systems (e.g. RobotReviewer) can fill data tables from full-texts. But these typically require annotated training data and careful rule-based development. The emergence of **generative AI and LLMs** is taking this further: modern LLMs need only natural-language instructions and can leverage knowledge learned during pre-training on huge corpora.

Generative LLMs (ChatGPT, GPT-4, etc.) were first applied widely around 2022–2023. They quickly demonstrated abilities to summarize text and answer questions, prompting researchers to explore them for literature tasks. A flurry of 2023–2025 studies now assesses their roles in evidence synthesis. For instance, an ISPOR Working Group (2025) explicitly identified three HTA areas for generative AI: systematic reviews, real-world data analysis, and health-economic modeling (<sup>[25]</sup> [www.ispor.org](#)). The plain-language summary of their report states:

“Generative AI, especially large language models, could significantly improve how evidence is generated for healthcare decisions... It can automate tasks like proposing search terms, screening abstracts, and extracting data in literature reviews; analyze large unstructured datasets (e.g. clinical notes) for RWE; and assist in developing and validating health economic models” (<sup>[25]</sup> [www.ispor.org](#)).

Thus, industry and academia view LLMs as potent new tools for evidence generation. The excitement must be tempered with caution: LLMs often have “black box” elements, can hallucinate incorrect facts, and raise validation issues (<sup>[26]</sup> [www.ispor.org](#)) (<sup>[8]</sup> [systematicreviewsjournal.biomedcentral.com](#)). Nevertheless, leading agencies are acknowledging this

trend. In 2024 NICE issued a **position statement** on AI in evidence generation, signaling that “evidence considered by NICE will be informed by AI methods” and stressing transparency and robustness ([www.nice.org.uk](http://www.nice.org.uk)). The FDA (Jan 2025) released draft guidance on AI in regulatory submissions, advocating a risk-based framework focusing on model credibility (<sup>[9]</sup> [www.fda.gov](http://www.fda.gov)). The EU’s AI Act and EMA’s reflection paper further underline that some AI use in drug lifecycle is expected (and regulated).

Overall, the historical context is clear: dossier processes have become more complex and data-heavy, driving a need for innovative methods. LLMs represent a transformative technology promising to accelerate evidence assembly. The remainder of this report delves into how they work, what evidence we have of their performance, and how they might reshape the pharma value dossier landscape.

## LLM Technologies and Mechanisms

### Understanding Large Language Models

Modern LLMs are deep learning models (often transformer-based) trained on **massive text corpora** (billions of words). They learn statistical patterns of language, enabling them to generate text given any prompt. Under the umbrella of *generative AI*, models like GPT-5.2, Claude, and others have demonstrated capabilities that approach or surpass human performance on many language tasks (<sup>[17]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)) (<sup>[27]</sup> [link.springer.com](http://link.springer.com)). For example, GPT-5.2 can write computer code, answer complex questions, or translate texts with fluency. Key properties relevant to evidence automation are:

- **Few-shot instruction following:** Given examples or detailed prompts, an LLM can perform tasks (e.g. generate Boolean search strings from PICO descriptions (<sup>[16]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)), extract data points from text) without task-specific training.
- **Knowledge integration:** LLMs (especially general ones) have broad world knowledge, including biomedical facts, gleaned from training text. However, their knowledge only extends until a training cutoff (for GPT-4, roughly late 2023; current models like GPT-5.2 have more recent cutoffs).
- **Need for guidance:** LLMs do best when guided by *engineering prompts*. Recent studies use techniques like chain-of-thought or retrieval augmentation (feeding them the actual text of papers) to improve accuracy (<sup>[28]</sup> [journals.plos.org](http://journals.plos.org)) (<sup>[7]</sup> [journals.plos.org](http://journals.plos.org)).
- **Rapid iteration:** LLMs can process text orders-of-magnitude faster than humans, constrained mainly by API rate limits (<sup>[4]</sup> [journals.plos.org](http://journals.plos.org)).

Beneath the hood, LLMs predict the next word token, but the emergent behavior is to simulate “reasoning” by leveraging vast associations learned during training (<sup>[17]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)) (<sup>[27]</sup> [link.springer.com](http://link.springer.com)). Several domain-specific variants are emerging (e.g. BioGPT, MedPaLM) that may better understand medical jargon. But even vanilla GPT models have proven surprisingly adept in health applications.

### Integration with Retrieval (RAG) and Pipelines

A crucial tactic in applying LLMs to evidence collection is *Retrieval-Augmented Generation (RAG)*. Instead of relying on the LLM’s static knowledge, RAG frameworks first retrieve relevant document content (e.g. abstracts from PubMed) and present that to the model. The LLM then answers questions or summarizes based on the provided evidence (<sup>[7]</sup> [journals.plos.org](http://journals.plos.org)). This approach was used in the AI-LES pipeline: the script supplied each article’s text to ChatGPT, significantly reducing “hallucinations” and ensuring answers were grounded in the actual content (<sup>[29]</sup> [journals.plos.org](http://journals.plos.org)) (<sup>[7]</sup> [journals.plos.org](http://journals.plos.org)).

By iteratively querying the LLM with references to specific sections or tables, systems can mimic a human reviewer's reading. For example, a retrieval-assisted prompt might include an abstract, and the model is asked: "List the study's primary outcome results from this abstract." The LLM then directly quotes or paraphrases from the text. Combining LLMs with database APIs (PubMed, Embase) or search tools (Semantic Scholar, etc.) enables end-to-end pipelines that discover and parse literature with minimal human scripting.

## Challenges of LLM Use

Despite their power, LLMs have notable limitations. Their training on broad internet text can introduce biases or outdated knowledge. In evidence synthesis:

- **Hallucinations:** Models sometimes fabricate plausible-sounding but false statements, especially when the real answer is unclear (<sup>[7]</sup> journals.plos.org) (<sup>[8]</sup> systematicreviewsjournal.biomedcentral.com). In evidence tasks, this can lead to incorrect data being reported. Developers mitigate this by feeding the model actual target text (RAG) and by designing prompts carefully (<sup>[28]</sup> journals.plos.org) (<sup>[7]</sup> journals.plos.org).
- **Lack of Domain Constraints:** LLMs may not inherently respect the rigorous structure of scientific data. They might confuse mean vs median, or extrapolate data trends without justification (<sup>[7]</sup> journals.plos.org). In critical steps, human validation is still required.
- **Transparency and Reproducibility:** Current LLMs are often "black box" neural nets. For HTA or regulatory use, agencies demand audit trails. Any LLM-assisted analysis needs logging of prompts and outputs, plus human checks ([www.nice.org.uk](http://www.nice.org.uk)) (<sup>[9]</sup> www.fda.gov).
- **Data Privacy:** Using patient-level data to train or prompt LLMs is generally not done (models are trained on text, not direct PHI). However, integrating unlabeled clinical notes could risk confidentiality and must comply with regulations.
- **API Constraints:** The cost and rate limits of commercial LLM APIs (ChatGPT, GPT-5.2) can be significant for large-scale reviews. Integrations with open-source models might help, but those may be less capable currently.

In sum, LLMs are potent but need governance. The breakthroughs in the next sections come from carefully constrained usage: human-guided pipelines where the model automates tedious tasks under expert supervision. With that in mind, we examine specific application areas.

## Automating Literature Searches

A first step in evidence collection is identifying relevant studies. Traditionally, a librarian or analyst crafts complex Boolean queries for databases like PubMed, Embase, Cochrane CENTRAL, etc. This demands expertise in medical subject headings (MeSH), synonyms, and trial designs. Even then, recall can be low: important studies may be missed due to wording variations or semantic gaps.

LLMs can assist by generating and refining search terms. TrialMind's approach exemplifies this: given a PICO description, GPT-4 produced candidate search keywords for interventions and conditions (<sup>[30]</sup> pmc.ncbi.nlm.nih.gov). Its interface allowed users to add or trim terms before running the query. In the experiments, GPT-generated queries recovered 71–83% of target studies (the "ground truth" set from published systematic reviews), whereas manually crafted queries by experts recovered only 13–23% (<sup>[1]</sup> pmc.ncbi.nlm.nih.gov). In particular, the ChatGPT/Lit approach could suggest synonyms or related terms that a human might overlook. The result was a much higher recall of relevant trials in the initial search. (Table 1 summarizes these improvements.)

More generally, studies have shown GPT-type models can excel in query formulation. Researchers have used chain-of-thought prompting to guide LLMs through PICO elements, producing effective queries (<sup>[16]</sup> pmc.ncbi.nlm.nih.gov) (<sup>[31]</sup> www.sciencedirect.com). For instance, an LLM can convert the plain-text review question ("adults with disease X should take drug Y vs standard therapy Y") into structured Boolean expressions automatically. One experiment for pandemic

drug discovery used GPT-4 to filter thousands of papers on SARS-CoV-2 and Nipah virus: the AI pipeline achieved high classification accuracy ( $\approx 92\%$  accuracy, F1 score  $\sim 88\%$  for SARS-CoV-2) by combining broad keyword search with LLM filtering (<sup>[32]</sup> [www.sciencedirect.com](http://www.sciencedirect.com)). This underscores that LLMs can effectively narrow down vast literatures once properly guided.

LLMs can also scan the retrieved records to identify which ones match eligibility criteria. In practice, the search step is often iterative: after initial screening, PICO may need refinement. LLMs can rapidly adjust queries based on inclusion/exclusion criteria changes. This agility helps when dossiers require living updates: as new questions arise (e.g. a secondary endpoint PICO), an LLM can generate a fresh query in minutes.

**Practical considerations:** LLM-based search automation should still use standard databases and follow established protocols (e.g. running parallel searches in PubMed and Embase). AI can augment, not replace, human search strategy. The NICE position also emphasizes that any tools “should only be used when there is demonstrable value” ([www.nice.org.uk](http://www.nice.org.uk)). In other words, companies should document how an LLM changed (and improved) search outputs compared to prior methods, to build trust with HTA reviewers.

## Automating Abstract and Full-Text Screening

Once papers are retrieved, the next step is screening titles/abstracts to identify studies meeting the review’s criteria. Traditionally this is done by two independent reviewers reading thousands of abstracts. Automating this step can save huge effort.

LLMs have shown remarkable performance here. In one study in radiology literature, researchers compared ChatGPT’s screening on 1,198 abstracts against that of human physicians (<sup>[2]</sup> [bmcmmedresmethodol.biomedcentral.com](http://bmcmmedresmethodol.biomedcentral.com)). ChatGPT completed the screening task in **1 hour**, whereas the doctors took **7–10 days** on average. In accuracy terms, ChatGPT achieved **95% sensitivity** (i.e. identified 95% of truly eligible papers) and an NPV (negative predictive value) of 99% (<sup>[2]</sup> [bmcmmedresmethodol.biomedcentral.com](http://bmcmmedresmethodol.biomedcentral.com)). It had very low false negatives, meaning it rarely missed an eligible study. Its specificity (true negatives) and PPV were lower than human reviewers, indicating more false positives, but in systematic reviews high sensitivity is often prioritized (missing a relevant trial is worse than including an extra one). The average agreement (kappa) between ChatGPT and human raters was modest (0.27) (<sup>[2]</sup> [bmcmmedresmethodol.biomedcentral.com](http://bmcmmedresmethodol.biomedcentral.com)), reflecting different error patterns, but overall the tool dramatically reduced workload by **40–83%** as calculated by the authors. Importantly, ChatGPT’s mistakes tended to be marginal cases. The model appears well-suited as a **first-pass filter**: it can quickly triage out clear negatives and flag likely inklings, after which humans need review only a smaller pool.

Similarly, TrialMind’s study screening module used GPT-4 to rank citations. They embedded known “ground truth” studies within a larger candidate set and had the model score each on inclusion likelihood. The resulting ranking brought the relevant studies much higher (fold change 1.5–2.6× over baseline methods) (<sup>[1]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)), meaning reviewers had to examine fewer abstracts to find the key ones. This kind of *prioritization* saves time. In the pilot, human–AI collaboration (where the AI did initial ranking and humans validated) boosted recall by 71.4% (<sup>[19]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)), compared to unguided screening.

Other groups confirm LLMs’ screening potential. Reviews and meta-scholar initiatives (e.g. ICASR 2023) noted LLMs can significantly assist screening but warned that errors must be checked (<sup>[33]</sup> [systematicreviewsjournal.biomedcentral.com](http://systematicreviewsjournal.biomedcentral.com)). Indeed, any automated screening should be followed by human verification of exclusions. Organizations are experimenting with *hybrid workflows* (LLM + human) to optimize both speed and accuracy.

In practice:

- An LLM-based screening tool might work as follows: It ingests each abstract (or full text) and is prompted with inclusion rules (stored as text prompts). It outputs a decision (“Include/Exclude”) and possibly a confidence score.

- Tools like ChatGPT can do this via the API. Alternatively, fine-tuned models (or classification tasks built on LLM embeddings) could be used.
- When uncertainty is high (LLM unsure), the item is sent to human reviewers. This is akin to triaging the easy cases.
- Importantly, all AI decisions should be traceable. For HTA use, NICE expects transparency about how AI was applied ([www.nice.org.uk](http://www.nice.org.uk)).

The bottom line: LLMs can drastically reduce the human workload in abstract screening with high sensitivity (<sup>[2]</sup> [bmcmcdresmethodol.biomedcentral.com](http://bmcmcdresmethodol.biomedcentral.com)). Table 1 (above) compares typical metrics.

## Automating Data Extraction and Synthesis

After screening, the accepted studies must be carefully mined for data points (e.g. sample sizes, outcome means, hazard ratios). This is often the most time-consuming manual step. LLMs can again help by reading scientific text and extracting structured results.

### Structured Data Extraction

AI-LES (PLOS One 2025) is a clear demonstration. The authors wrote a Python script (“AI-LES”) that took each paper’s PDF, passed the text to ChatGPT, and asked fixed questions (e.g. “What is the mean incubation period and its 95% CI?”). They tested on 94 COVID-19 epidemiology papers. For **clearly reported values**, ChatGPT was 100% accurate (<sup>[3]</sup> [journals.plos.org](http://journals.plos.org)). Crucially, AI-LES could **distinguish** key fields (means vs medians, CI vs IQR, etc.) with perfect accuracy when the paper explicitly stated them (<sup>[3]</sup> [journals.plos.org](http://journals.plos.org)). Errors occurred only when the paper’s language was ambiguous—just as a human often struggles with incomplete reporting (<sup>[7]</sup> [journals.plos.org](http://journals.plos.org)).

Even more impressively, AI-LES was **much faster**. It processed 94 articles in 76 minutes total (≈48 seconds/article) (<sup>[4]</sup> [journals.plos.org](http://journals.plos.org)). Of that, only ~11 seconds per paper was actual LLM processing; the remainder was waiting on the API. In comparison, manual extraction took about **7 times longer** for the same set (<sup>[4]</sup> [journals.plos.org](http://journals.plos.org)). The paper notes that AI-LES scaled well: abstracts or tables that slow down humans (e.g. complex PDF layouts) did not similarly hinder the AI.

TrialMind also tackled data extraction (beyond screening). It broke papers into defined fields (study protocols, baselines, outcomes) and asked GPT-4 to pull numbers. The system outperformed vanilla GPT-4 by 16–32% in accuracy (<sup>[1]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)), presumably due to optimized prompts and instructions. In human–AI trials, the hybrid method increased extraction accuracy +23.5% and cut time by 63.4% (<sup>[19]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)).

Another data point: Booth et al. (2024) used GPT-4 for network meta-analyses (meta-analyses requiring multiple treatments). Across four published NMAs, GPT-4 built R code and extracted data with >99% accuracy (<sup>[6]</sup> [link.springer.com](http://link.springer.com)). It output fully functioning analysis scripts directly from the literature, including tables and evergreen results. The authors emphasize that the LLM-required only “routine checks” afterwards (<sup>[34]</sup> [link.springer.com](http://link.springer.com)), achieving highly efficient synthesis.

Table 1 (below) summarizes empirical results from these case studies, showing that for data extraction tasks, LLM-augmented methods matched or exceeded human performance while significantly reducing time.

Table 1: Examples of LLM-assisted extraction outcomes

| Extraction Task                                | Human   | LLM-assisted               | Result  |
|--|---|----------------------------|---|
| Extraction of study parameters (e.g. mean, CI) | Manual review via Excel (532 min for 94 articles) | AI-LES (ChatGPT) via API   | 100% accuracy when reported clearly ( <sup>[3]</sup> <a href="http://journals.plos.org">journals.plos.org</a> ); 7× faster overall (76 min vs 532 min) ( <sup>[4]</sup> <a href="http://journals.plos.org">journals.plos.org</a> )                                      |
| Extraction of trial outcomes                   | Two analysts (long hours per study)               | GPT-4 pipeline (TrialMind) | AI+human improved accuracy by 23.5%; screening time –63% ( <sup>[19]</sup> <a href="http://pmc.ncbi.nlm.nih.gov">pmc.ncbi.nlm.nih.gov</a> ); expert pref. TrialMind in 62–100% cases ( <sup>[19]</sup> <a href="http://pmc.ncbi.nlm.nih.gov">pmc.ncbi.nlm.nih.gov</a> ) |

| Extraction Task              | Human                                     | LLM-assisted                                | Result  |
|------------------------------|---|---|---|
| NMA data & code extraction   | Statistician codes R scripts manually     | GPT-4 (PaLM4All) generates code and results | >99% accurate data pull ( <sup>[6]</sup> link.springer.com); valid R scripts end-to-end; correct interpretation ( <sup>[6]</sup> link.springer.com)                                       |
| Health econ model parameters | Health economist programs Excel/R scripts | GPT-4 reads design text and outputs code    | Accurate model replication: 100% (14/15 runs) for NSCLC model ( <sup>[35]</sup> pmc.ncbi.nlm.nih.gov); 93% fully error-free; 60–87% for RCC model ( <sup>[35]</sup> pmc.ncbi.nlm.nih.gov) |

These results suggest LLMs can take over much of the **data trawling** in evidence synthesis. For clearly reported numbers, AI is essentially foolproof (<sup>[3]</sup> journals.plos.org). The task of converting narrative results (e.g. “mean OS = 12.7 months, 95% CI 9.8–15.6”) into structured numbers is exactly what LLMs were trained to do when prompted correctly.

## Unstructured Data Sources

In addition to published literature, LLMs could be used to extract evidence from unstructured sources such as clinical notes or registries; this remains nascent. The ISPOR report notes that LLMs can analyze “unstructured clinical notes and imaging” in RWE (<sup>[25]</sup> www.ispor.org). For example, a pharmaceutical company might use NLP on EHR data to identify patient outcomes or adverse events patterns. While generic LLMs are not trained on private patient data, one could use them on de-identified real-world text. Caution is high due to privacy. So far, most evidence collection is from **public text** (journals, databases).

## Summarization and Narrative Generation

Beyond numeric extraction, LLMs can help *draft* the narrative sections of a dossier (the value story). Companies often need to write persuasive but evidence-based text summarizing clinical benefits, burden of illness, and economic impact. Early tests show ChatGPT can generate reasonably coherent summaries of trial results when given data (<sup>[18]</sup> pubmed.ncbi.nlm.nih.gov) (<sup>[36]</sup> link.springer.com), but factual precision is spotty. For example, an AI-generated summary might weave data points into a narrative, but it requires careful editing to ensure no misinterpretation. Some groups are exploring “AI-assisted writing”: the LLM produces an initial draft which experts refine. The variations in human writing style and regulatory tone suggest caution: any LLM-generated text would have to be thoroughly validated against sources (<sup>[8]</sup> systematicreviewsjournal.biomedcentral.com) ([www.nice.org.uk](http://www.nice.org.uk)).

However, given the current results in extraction and analysis, LLMs can at least supply factual bullet points for value messages. For instance, after mining all trials, an AI could be prompted: “Based on the following data... write a 200-word summary of efficacy results.” Its output could then be edited. This is still a frontier area, with limited published studies. But it is a logical next step in using LLMs to automate dossier content generation.

## Automated Health Economic Modeling

Clinical evidence feeds directly into economic models (e.g. Markov models, partitioned survival models). Automating economic modeling has historically been even less explored; yet it is part of the “value evidence” chain. Recent work by Reason et al. (2024) and colleagues has tested GPT-4’s ability to actually program economic models.

In *PharmacoEconomics – Open*, Reason et al. conducted a **case study** to see if GPT-4 could reproduce published cost-effectiveness models from textual specifications (<sup>[37]</sup> pmc.ncbi.nlm.nih.gov) (<sup>[5]</sup> pmc.ncbi.nlm.nih.gov). They described two three-state partitioned-survival models (for lung and renal cancer treatments), gave GPT-4 the design, assumptions, and parameters, and asked it to write R code. The results were striking: GPT-4 **fully replicated** the lung cancer model 100% of the time (15/15 trials), with 93% of scripts entirely error-free (<sup>[5]</sup> pmc.ncbi.nlm.nih.gov). For the second model, after simplifying a complex calculation, GPT-4 produced 87% scripts with ≤1 minor error and 60% 100% correct (<sup>[5]</sup> pmc.ncbi.nlm.nih.gov). The outputs replicated the published incremental cost-effectiveness ratios within 1%.

This indicates that LLMs can handle the programming logic and numeric precision needed for economic models. The **Key Points** of that study conclude that GPT-4 has practical potential in automating model construction ([38] pmc.ncbi.nlm.nih.gov), which could greatly shorten development times and reduce manual coding errors. (Indeed, human-built models are known to often contain logic bugs ([39] pmc.ncbi.nlm.nih.gov).)

Another group (Booth et al., *PharmacoEconomics – Open* 2024) focused on network meta-analyses (NMA), which feed into health economic models. As noted earlier, they had GPT-4 extract data and generate R scripts for multiple NMAs with >99% accuracy ([6] link.springer.com). This shows that even complex statistical tasks (like writing forest plots) can be proposed by LLMs.

The implications are that the **entire pipeline** of HTA (from evidence search to final model outcome) could be integrated into an LLM-augmented workflow. The model and data extraction could be done in tandem: an LLM finds all trial arms, outcomes, and then automatically estimates life years, QALYs, costs, and even sensitivity analyses. Regulatory guidance is nascent here, but an ISPOR report notes that while such tools are at an early stage, they recommend continued evaluation ([10] pubmed.ncbi.nlm.nih.gov).

In summary, LLMs have demonstrated the capability to **accelerate health economics**: one can prompt an LLM like GPT-5.2 to generate modeling code and results, then have experts check and refine. This could cut months of Excel/R development time to hours of prompting and debugging.

## Case Studies and Evidence Summary

To illustrate the range of LLM applications, Table 2 below summarizes notable published case studies. Each integrates LLMs in different parts of the evidence generation chain.

| Study (reference)  | Process / Task                       | LLM & Approach                 | Key Results (LLM vs Traditional)   |
|--|--------------------------------------|--------------------------------|--|
| Wang et al., <i>NPJ Digit Med</i> 2025 ([40] pmc.ncbi.nlm.nih.gov)                           | Systematic Review Search & Screening | GPT-4 (TrialMind pipeline)     | Recalls 0.711–0.834 vs 0.138–0.232 (human) ([41] pmc.ncbi.nlm.nih.gov); screening ranking +1.5–2.6x ([41] pmc.ncbi.nlm.nih.gov); combined AI+human recall +71.4%, screening time +44.2% ([19] pmc.ncbi.nlm.nih.gov). |
| DeOliveira et al., <i>BMC Med Res Method</i> 2024 ([2] bmcmredresmethodol.biomedcentral.com) | Abstract Screening                   | ChatGPT (GPT-3.5)              | Sensitivity 95%, NPV 99% ([2] bmcmredresmethodol.biomedcentral.com); time: 1198 abstracts in 1h (ChatGPT) vs 7–10 days (humans); workload saved 40–83%.  |
| Mitchell et al., <i>PLOS One</i> 2025 ([3] journals.plos.org) ([4] journals.plos.org)        | Data Extraction (key values)         | ChatGPT (AI-LES Python script) | Accuracy 100% (for clear stats) ([3] journals.plos.org); 94 articles in 76 min (AI) vs ~532 min (human) ([4] journals.plos.org); ~7x speedup.  |
| Booth et al., <i>PharmacoEconomics</i> 2024 ([6] link.springer.com)                          | Network Meta-Analysis synthesis      | GPT-4 via API                  | >99% of data correctly extracted over 4 NMAs ([42] link.springer.com); generated functioning R scripts; accurate interpretation and reporting.   |
| Reason et al., <i>PharmacoEconomics</i> 2024 ([5] pmc.ncbi.nlm.nih.gov)                      | Health Econ Model Generation         | GPT-4 (text prompts → R code)  | NSCLC model: 100% replicated (14/15 runs), 93% error-free ([35] pmc.ncbi.nlm.nih.gov); RCC model: 87% replicated (9/15 fully correct) ([35] pmc.ncbi.nlm.nih.gov); ICERs within 1%.                                  |
| Liu et al., <i>Int. J. Med. Inform</i> 2024 ([32] www.sciencedirect.com)                     | Rapid Lit. Review for Drug Targets   | GPT-4 (pipeline)               | SARS-CoV-2: Acc 92.9%, F1 88.4% ([32] www.sciencedirect.com); Nipah: Acc 87.4%, F1 73.9%. (Automated pipeline closely matched human expert labeling.)  |

Each case study spans a different domain: cancer vs infectious disease, structured vs unstructured data, etc. They consistently show that GPT models can either match or dramatically accelerate human processes with high fidelity. For instance, the NPJ study found the **AI-synthesizer was preferred by experts** over naïve GPT output ([19] pmc.ncbi.nlm.nih.gov). The PLOS and BMC studies quantify speedups of 5–7x. The PharmacoEcon studies (Reason and Booth) demonstrate high **reproducibility** of results. These examples give empirical weight to the proposition that LLMs can **automate crucial evidence tasks**.

## Perspectives of HTA Bodies and Regulators

While the technology advances, acceptance by payers and regulators is paramount. HTA agencies traditionally emphasize transparency, rigor, and replicability ([www.nice.org.uk](http://www.nice.org.uk)) (<sup>[33]</sup> [systematicreviewsjournal.biomedcentral.com](https://systematicreviewsjournal.biomedcentral.com)). How do these bodies view AI-generated evidence?

- **NICE (UK):** In 2024 NICE published a *position statement* on AI in evidence generation ([www.nice.org.uk](http://www.nice.org.uk)). Major points: AI offers potential “superior approaches” by processing large data efficiently ([www.nice.org.uk](http://www.nice.org.uk)), but concerns exist over transparency and trustworthiness. NICE expects that any AI use be *declared, transparent, reproducible* and that human oversight remains central. AI-derived results must meet the same quality standards as conventional methods. NICE encourages early discussion on AI use in submissions ([www.nice.org.uk](http://www.nice.org.uk)). This suggests HTA bodies will allow AI-augmented methods, provided sponsors thoroughly document them.
- **EMA (Europe):** The European Medicines Agency (EMA) has a 2024 “reflection paper” on AI (published Oct 2024 (<sup>[43]</sup> [www.putassoc.com](http://www.putassoc.com))). It outlines principles for AI in drug lifecycle, focusing on context, validation, and risk-based approaches. EMA itself pilots ML for literature search. (<sup>[44]</sup> [www.putassoc.com](http://www.putassoc.com)) EMA is coordinating an EU “AI strategy in medicines regulation” and training staff. Though not HTA, this signals EU regulators are preparing for AI-run analysis, implying dossiers could increasingly cite AI methods.
- **FDA (US):** In January 2025 the FDA issued draft guidance for AI use in drug submissions (<sup>[9]</sup> [www.fda.gov](http://www.fda.gov)). It stresses defining context-of-use and establishing model credibility (through testing, comparators, bias analysis). The FDA encourages sponsors to engage the agency early on AI credibility assessment (<sup>[9]</sup> [www.fda.gov](http://www.fda.gov)). This would cover, for example, defending an LLM-derived analysis package. Crucially, it expects “agile, risk-based” frameworks that still ensure scientific standards (<sup>[45]</sup> [www.fda.gov](http://www.fda.gov)) (<sup>[9]</sup> [www.fda.gov](http://www.fda.gov)).
- **Other HTA Agencies:** Most national HTA bodies have not yet issued detailed AI guidelines. A 2025 survey by Putnam noted that aside from NICE, countries from Italy to Australia have no specific HTA AI standards (<sup>[46]</sup> [www.putassoc.com](http://www.putassoc.com)). Germany’s IQWiG allows ML for search but has no formal policy. France’s HAS is testing AI for literature reviews but wary of limitations (<sup>[47]</sup> [www.putassoc.com](http://www.putassoc.com)). Canada’s CADTH is monitoring developments. In sum, HTA policy is just forming. Agencies are aware AI is “happening” (<sup>[48]</sup> [www.putassoc.com](http://www.putassoc.com)) but emphasize the analogy with Real-World Evidence: initially skeptical, now gradually incorporating standards.

Taken together, these statements show support for AI’s promise, but under strong conditions. HTA reviewers will likely require:

- **Declaration of AI use:** Dossier sections should specify if AI tools assisted (e.g. “search strategy was generated with AI, model: GPT-5.2”).
- **Justification & transparency:** How was the tool used? Which steps? Provide enough detail for replication (e.g. prompt logs, code).
- **Human oversight:** Verify AI output at each stage; possibly dual submission (AI-assisted draft + expert annotation).
- **Validation:** Demonstrate that AI methods produce results comparable to accepted standards (for example, pilot runs comparing human vs AI on sample data).
- **Sensitivity analyses:** Explore robustness to AI errors (for example, re-running analyses with any alternative data).

The **challenge** is that LLMs evolve quickly. Agencies may take a cautious “trust but verify” stance (<sup>[49]</sup> [www.putassoc.com](http://www.putassoc.com)) (<sup>[8]</sup> [systematicreviewsjournal.biomedcentral.com](https://systematicreviewsjournal.biomedcentral.com)). They draw parallels to the earlier debate on RWE: initial guidance was minimal, but over time consensus methods emerged. An ISPOR report recommends pilot projects where manufacturers and HTA bodies co-develop AI pipelines, to build confidence (<sup>[50]</sup> [www.putassoc.com](http://www.putassoc.com)).

Even beyond HTA, **ethics and equity** are rising concerns. Model biases or missing data transparency could affect vulnerable populations differently. The NICE statement specifically warns of algorithmic bias and patient privacy ([www.nice.org.uk](http://www.nice.org.uk)). Sponsors must ensure that using AI does not compromise data integrity.

Regulatory and HTA acceptance of AI-derived evidence is in its infancy. But the trend is clear: “While use in submissions remains very limited to date, ongoing developments... signal a shift toward more structured adoption” (<sup>[51]</sup> [www.putassoc.com](http://www.putassoc.com)). In other words, LLMs will likely become standard tools in dossier preparation, but under regulated frameworks that demand rigorous documentation and validation.

## Implications and Future Directions

The integration of LLMs into evidence generation is a fast-moving frontier. The studies cited above provide early validation of feasibility. Looking forward, several implications and future paths emerge:

- Comprehensive Pipelines:** We can envision **end-to-end AI systems** for HTA evidence. A single workflow might take trial identifiers or molecular targets as input, retrieve all publications, screen them, extract numeric outcomes, incorporate RWE, and even build preliminary health-economic models – all with minimal human input. Platforms like “TrialReviewBench” (from TrialMind) lay groundwork for this. Additional innovations like RAG databases and specialized LLM interfaces (e.g. [Elicit.ai](#) or [SciBERT](#)) will make pipelines more robust.
- Living Dossier Updates:** HTA rarely stays static; new data emerge. LLMs naturally enable “living” systematic reviews: the AI script could regularly poll for new trials and incorporate them into the model. Indeed, Mitchell et al. highlight that AI tools lower barriers to keeping reviews up-to-date (<sup>[52]</sup> [journals.plos.org](#)) (<sup>[53]</sup> [journals.plos.org](#)). In future, a value dossier could be dynamically updated by alerts rather than entirely rewritten before resubmission.
- Specialized LLMs:** General LLMs are good, but domain-specific models are forthcoming (e.g. BioGPT2, Med-PaLM2, etc.). These may further improve performance on niche tasks (e.g. interpreting pathology imaging). In section analysis, combining vision+language models could let an AI “read” scanned charts or imaging findings too.
- Human–AI Team:** The best practice is *hybrid*. The consensus of many authors is that AI will augment, not replace, experts (<sup>[33]</sup> [systematicreviewsjournal.biomedcentral.com](#)) (<sup>[54]</sup> [www.putassoc.com](#)). LLMs handle bulk work, flag anomalies, and summarize. Experts provide the domain judgment, catch AI errors, and weave the final narrative. This human–machine teaming is likely to become a standard part of pharmaceutical R&D workflows.
- Regulatory Evolution:** HTA and regulatory bodies will catch up. We expect new guidelines specifically covering AI tools in evidence synthesis. For instance, the EU AI Act classifies tools by risk – an LLM that outputs medical evidence might be “high-risk”, requiring stricter controls. Journals and conferences will start demanding authors disclose AI usage too (as evidenced by ISPOR’s plain-language note (<sup>[55]</sup> [www.ispor.org](#)) and the general editorial policies evolving). Firms will need new Standard Operating Procedures (SOPs) for audited AI processes.
- Data and IP Issues:** One practical issue is that many full-text papers are behind paywalls; feeding their text into closed LLM APIs might conflict with publishers’ terms. Solutions could involve university/library subscriptions or retraining open models on legal corpora. Meanwhile, specialized LLM hosting (on-premise) might be needed for sensitive data.
- Ethical Considerations:** As AI do more of the writing, we run into questions of authorship and accountability. If an analysis is 80% machine-generated, who certifies it? The field will grapple with how to credit AI’s role vs. human authorship. Also, addressing bias in training data (LLMs trained on older literature may underrepresent certain populations or treatments) will be critical.
- Business Impact:** For industry, mastering AI in evidence generation could become a competitive advantage. Companies that harness these tools can prepare dossiers faster, possibly reaching markets sooner. It could shift the R&D timeline – the bottleneck of evidence synthesis is eased. In the long term, this might lead to **smaller, faster trials** too, if LLM analysis can squeeze more from less data.

Table 3 (below) projects a **future workflow** scenario for an evidence dossier using LLMs:

| Stage                     | Traditional Approach                                  | LLM-augmented Future  |
|---------------------------|---|---|
| Formulating PICO          | Experts brainstorm keywords and logic gates manually. | LLM suggests comprehensive search strings from plain query.   |
| Literature Search         | Run database queries, screen titles by hand.          | Run AI-enhanced search; LLM filters by relevance automatically.   |
| Abstract Screening        | Two reviewers read ~thousands of abstracts (weeks).   | LLM screens majority at high sensitivity within hours ( <sup>[2]</sup> <a href="#">bmcmcdresmethodol.biomedcentral.com</a> ); humans review edge cases. |
| Data Extraction           | Analysts open PDFs, manually record outcomes.         | LLM parses papers (via API) and auto-populates extraction tables ( <sup>[3]</sup> <a href="#">journals.plos.org</a> ).                                  |
| Synthesis (Meta-analysis) | Statistician codes R/Excel, runs models.              | LLM generates analysis code and figures ( <sup>[6]</sup> <a href="#">link.springer.com</a> ). Human reviews outputs.                                    |
| Modeling                  | Health economist codes model (months).                | LLM codes model from text spec ( <sup>[5]</sup> <a href="#">pmc.ncbi.nlm.nih.gov</a> ); performs calibration.   |
| Report Writing            | Team writes narrative, including evidence tables.     | LLM drafts sections of the value story, which staff refine for accuracy.  |

| Stage                       | Traditional Approach  | LLM-augmented Future   |
|-----------------------------|---|--|
| Review/Regulator Engagement | Experts justify every step, often prodding for underlying data. | Documentation of LLM prompts and checks ensures transparency; experts explain any AI anomalies per [NICE guidelines] ( <a href="http://www.nice.org.uk">www.nice.org.uk</a> ). |

Table 3: Illustrative future workflow using LLM automation (based on literature findings).

This depiction is speculative but grounded in current trials of technology. The transition to such workflows will likely be gradual and uneven. Early adopters might use LLMs internally for scoping or preliminary analyses, while full AI-driven submissions will await regulatory comfort.

## Conclusion

The integration of large language models into the evidence generation pipeline for economic dossiers and value stories is already underway and looks set to accelerate. Recent published studies demonstrate that GPT-based systems can **match or exceed human expert performance** in literature search, screening, data extraction, and even model building – and can do so far faster. For example, GPT-4 outperformed traditional search recall by ~3× (<sup>[41]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)) and replicated complex economic models with >90% correctness (<sup>[5]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)). Industry and HTA thought leaders recognize this potential: an ISPOR working group titled “Generative AI for HTA” (2025) confirms the promise of AI-driven systematic reviews and economic modeling (<sup>[25]</sup> [www.ispor.org](http://www.ispor.org)).

Helping drive this acceptance is the evidence that LLMs can become trusted assistants rather than mysterious black boxes. Hybrid workflows, where AI outputs are carefully validated, can offer the “best of both worlds”: massive efficiency gains without sacrificing scientific rigor. The case study data compiled above show experts pre-validating AI pipelines can save months of work. As one author concluded, GPT-4’s assistance could accelerate model development timelines and *reduce the risk of error* (<sup>[38]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)).

The future implications are profound. When new therapies are evaluated, we may see “**living value dossiers**” that update in real time with new data, AI-powered narrative updates, and dynamic economic forecasts. This could shorten time to patient access—an outcome explicitly valued by payers and the public. At the same time, careful policy and practice safeguards must be established to ensure trust. Stakeholders should collaborate (industry, HTA, regulators, technologists) to share methods, validate tools on reference datasets, and create standards.

In summary, LLMs are poised to revolutionize how clinical evidence is collected and synthesized for pharmaceutical value communication. They promise a leap in efficiency, consistency, and possibly creativity in constructing value propositions. But Prudence is required: as NICE and others caution, AI is a *tool*, not a replacement for scientific judgment ([www.nice.org.uk](http://www.nice.org.uk)). The path forward will involve rigorous testing, transparency, and phased adoption of AI. If done correctly, AI-enhanced evidence generation could become a new “standard of care” in HTA submissions, enabling faster, data-rich decision-making that ultimately benefits patients and health systems.

**References:** This report draws on recent research and expert reports. Key references include ISPOR guidance (<sup>[25]</sup> [www.ispor.org](http://www.ispor.org)) (<sup>[56]</sup> [pubmed.ncbi.nlm.nih.gov](http://pubmed.ncbi.nlm.nih.gov)), NICE’s position statement ([www.nice.org.uk](http://www.nice.org.uk)), trial automation studies (<sup>[1]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)) (<sup>[5]</sup> [pmc.ncbi.nlm.nih.gov](http://pmc.ncbi.nlm.nih.gov)) (<sup>[57]</sup> [journals.plos.org](http://journals.plos.org)), and several systematic reviews of AI in evidence synthesis (<sup>[2]</sup> [bmcmredsmethodol.biomedcentral.com](http://bmcmredsmethodol.biomedcentral.com)) (<sup>[3]</sup> [journals.plos.org](http://journals.plos.org)). All claims above are supported by peer-reviewed sources (citations in text).

## External Sources

[1] <https://pmc.ncbi.nlm.nih.gov/articles/PMC12331930/#:~:For%2...>





## IntuitionLabs - Industry Leadership & Services

**North America's #1 AI Software Development Firm for Pharmaceutical & Biotech:** IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

**Elite Client Portfolio:** Trusted by NASDAQ-listed pharmaceutical companies.

**Regulatory Excellence:** Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

**Founder Excellence:** Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

**Custom AI Software Development:** Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

**Private AI Infrastructure:** Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

**Document Processing Systems:** Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

**Custom CRM Development:** Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

**AI Chatbot Development:** Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

**Custom ERP Development:** Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

**Big Data & Analytics:** Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

**Dashboard & Visualization:** Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

**AI Consulting & Training:** Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at <https://intuitionlabs.ai/contact> for a consultation.

---

## DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by [Adrien Laurent](#), a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

© 2025 IntuitionLabs.ai. All rights reserved.