HTAi™
HEALTH TECHNOLOGY ASSESSMENT iNTERNATIONAL

GENERATIVE AI GLOSSARY FOR HTA PROFESSIONALS

A reference for HEOR practitioners working on the HTA side — submissions, evidence reviews, modelling, and methods.

Generative AI Glossary for HTA Professionals

A reference for HEOR practitioners working on the HTA side — submissions, evidence reviews, modelling, and methods.

Core concepts

Term	Definition	Relevance to HTA
Generative AI (GenAI)	A class of AI that produces new content (text, code, structured data, images) rather than only classifying or predicting.	Drafting submissions, summarising clinical evidence, restructuring extracted data, etc can be applied anywhere across HTA.
Large Language Model (LLM)	A neural network trained on large text corpora that generates language by predicting the next token. Examples: Claude, GPT, Gemini, Llama.	The engine behind most current tools used for AI in HTA
Foundation model	A large, broadly pre-trained model that can be adapted to many downstream tasks. LLMs are the most common type.	See above but with broader applications
Transformer	The neural network architecture underlying nearly all modern LLMs (introduced 2017).	Mostly background, but it explains why context length and “attention” costs scale the way they do.
Token	A sub-word unit of text (roughly three-quarters of an English word). Pricing, context limits, and latency are measured in tokens.	A single NICE STA submission, CSR or technical model report easily runs into the hundreds of thousands of pages and tokens which is relevant when scoping automation costs.
Context window	The maximum number of tokens (input + output) a model can consider at once.	Determines whether you can fit a full SLR PDF, a model technical document, or a comparator HTA report into one prompt.
Embedding	A numerical vector representing the semantic meaning of a chunk of text; similar meanings produce similar vectors.	Used to find conceptually related abstracts during SLR screening, or to match outcome definitions across studies that use different wording.
Vector database	A database optimised for storing and searching embeddings (e.g., Pinecone, Weaviate, pgvector).	The back-end for “search my evidence library” tools across HTA decisions, protocols, or model templates.

Generative AI Glossary for HTA Professionals

A reference for HEOR practitioners working on the HTA side — submissions, evidence reviews, modelling, and methods.

Building and training models

Term	Definition	Relevance to HTA
Pre-training	Initial training on large unlabeled corpora to learn general language patterns. Expensive, infrequent.	The reason LLMs are general purpose and can be applied across the HTA spectrum of work as well as to write poetry etc
Fine-tuning	Additional training of a pre-trained model on a smaller, curated dataset to specialise it.	A vendor might fine-tune on HTA submissions to improve tone, formatting or general quality of the work
RLHF (Reinforcement Learning from Human Feedback)	Training technique where humans rank outputs and the model learns to prefer favoured ones. Underpins most chat-tuned frontier models.	Explains why frontier models hedge, refuse certain prompts, and adopt particular tones.
Distillation	Training a smaller, cheaper model to mimic a larger one.	Relevant when choosing cost-effective deployment for things like screening if accuracy is not traded off too much
Knowledge cutoff	The date beyond which the model has no built-in knowledge of world events.	A model with a 2024 cutoff won't know about a 2025 NICE methods update or recent CADTH recommendations.
Open-weight / open-source model	A model whose weights (and sometimes training code) are publicly downloadable (e.g., Llama, Mistral).	Attractive for on-premises deployment so confidential clinical data never leaves institutional infrastructure.
Closed / proprietary model	A model accessible only via a vendor API (e.g., Claude, GPT-4 family, Gemini).	Usually higher capability, but requires careful governance and handling of confidential daa.

Interacting with models

Term	Definition	Relevance to HTA
Chatbot	A conversational interface (e.g., Claude.ai, ChatGPT, Gemini, Copilot) that wraps an LLM in a chat experience for end users.	The most accessible entry point for analysts, good for exploratory queries, summarising open-access papers, and drafting. Generally unsuitable for confidential trial data unless using an enterprise tier with appropriate data

Generative AI Glossary for HTA Professionals

A reference for HEOR practitioners working on the HTA side — submissions, evidence reviews, modelling, and methods.

Term	Definition	Relevance to HTA
		agreements; outputs are not automatically logged or reproducible.
API call	A programmatic request sent to a model endpoint (cloud or self-hosted) that returns a model output. The atomic unit of any automated GenAI workflow.	Enables batch processing of abstracts, structured data extraction at scale, and integration with R, Python, or Excel pipelines and building of software with LLMs. Each call is metered, loggable, and versionable, important for reproducible HTA work.
Prompt	The input text given to the model.	The difference between a poor and a well-structured prompt often can determine whether you get good outputs, though models are increasingly using auto prompting and producing good outputs from basic prompts
Prompt engineering	The practice of crafting prompts to reliably elicit useful outputs.	Of high importance to produce good outputs originally but perhaps becoming less relevant (see above)
System prompt	A higher-priority instruction that frames the model's role and behaviour for a session.	E.g., “You are a HEOR analyst extracting outcomes data suitable for a network meta-analysis from clinical trial publications”
Zero-shot	Asking the model to perform a task with no examples.	Fine for general queries; usually insufficient for nuanced extraction or appraisal tasks.
Few-shot	Providing a handful of input–output examples in the prompt.	Often needed for consistent extraction across heterogeneous publications and conference abstracts.
Chain-of-thought (CoT)	Prompting the model to reason step-by-step before answering.	Potentially useful for more complex tasks but as mentioned perhaps becoming less relevant
Temperature	A parameter (typically 0–2) controlling output randomness. Lower = more deterministic; higher = more variable.	Arguably should always be set to 0 for HTA due to reproducibility concerns

Beyond basic chat

Term	Definition	Relevance to HTA
Retrieval-Augmented Generation (RAG)	A pattern where the model retrieves relevant documents from an external store and uses them to ground its answer.	Used for very large corpuses of documents (I.e. many CSRs, HTA dossiers etc) to improve speed, accuracy and reduce hallucinations.
Tool use / function calling	The model's ability to call external functions (calculators, APIs, databases) during a response.	Enables hybrid and agentic workflows, e.g., the models formulate a query, runs a PubMed search, parses the results, and summarises them. Different models can be given access to different tools to execute their specific HTA specific tasks.
Orchestration	Coordinating multiple model calls, tools, and data sources into a coherent multi-step workflow. Common frameworks include LangChain, LlamaIndex, Semantic Kernel, Haystack, and low-code platforms like n8n.	Underpins HTA pipelines such as "search → de-duplicate → screen → extract → quality-assess → synthesise," where outputs from each stage feed the next. The orchestration layer is where validation, logging, and human checkpoints get implemented.
Agent / agentic workflow	A system where the model plans and executes multi-step tasks autonomously, typically using tools. A common output of an orchestration framework.	Emerging in automated SLR pipelines with human checkpoints between stages. Greater autonomy raises validation and audit demands.
Multimodal	A model that handles inputs or outputs beyond text — images, PDFs, audio, video.	Useful for parsing figures, Kaplan–Meier curves, forest plots, and scanned HTA documents.
Structured output / JSON mode	Constraining the model to return data in a defined schema.	Important for piping extracted study characteristics directly into Excel, R, or downstream analytics without manual reformatting.
Human-in-the-loop (HITL)	Workflow design where humans review or approve model outputs at defined checkpoints.	Baseline expectation for any output influencing a submission or methods decision.

Output quality and evaluation

Term	Definition	Relevance to HTA
Hallucination	When a model generates plausible-sounding but factually wrong or fabricated content including invented citations, made-up trial results, or non-existent TA numbers.	The big risk with LLMs. Mitigated, never eliminated, by RAG, structured outputs, and HITL.
Bias	Systematic skew in outputs arising from training data, RLHF choices, or prompt design.	Relevant for equity considerations in HTA and fair decision making.
Reproducibility	Whether the same prompt yields the same output. LLMs are stochastic by default.	A methodological challenge for pipelines feeding into submissions since reproducibility is a cornerstone of HTA. Can be mitigated with proper AI use and disclosing information about how AI outputs were generated.
Validation	Empirical demonstration that an AI tool performs adequately for a defined task on representative data, with documented sensitivity, specificity, or equivalent metrics.	Likely to be increasingly expected by HTA bodies before AI-derived evidence is accepted in submissions.
Benchmark / evaluation set	A standardised dataset and scoring rubric for comparing models.	Domain-specific benchmarks (e.g., for SLR screening recall) may start to emerge to validate specific models and workflows before allowing use in HTA.
Drift	Degradation of performance over time as model versions, data, or upstream context change.	A governance and accuracy issue, a tool validated last year may behave differently this year.
Explainability / interpretability	The extent to which a model's reasoning can be inspected.	Limited for LLMs; usually replaced operationally by source citations, prompt logs, and human review.

Governance, safety, and emerging policy

Term	Definition	Relevance to HTA
Guardrails	Rules or filters that constrain model behaviour (refusing harmful content, redacting PII, blocking off-topic outputs).	Essential for ensuring accurate and appropriate outputs for HTA and proper governance against, for example misuse of AI.

Generative AI Glossary for HTA Professionals

A reference for HEOR practitioners working on the HTA side — submissions, evidence reviews, modelling, and methods.

Term	Definition	Relevance to HTA
EU AI Act / regulatory frameworks	The EU AI Act and emerging national rules on AI in healthcare and medical devices.	GenAI tools embedded in clinical decisions can fall into “high-risk” categories with conformity-assessment obligations; tools supporting evidence generation sit on a more uncertain footing.
Responsible AI / AI governance	Organisational practices for safe, ethical, and accountable AI use covering procurement, validation, audit, and decommissioning.	We may increasingly be seeing specific frameworks from HTA bodies and professional societies as the field evolves.

GENERATIVE AI GLOSSARY FOR HTA PROFESSIONALS

Contributors:

- Tim Reason (UK), Co-Lead of the Coordination Group of the HTAi Community of Practice on AI for HTA activities.
- Seye Abogunrin (Switzerland), Co-Lead of the Coordination Group of the HTAi Community of Practice on AI for HTA activities

Acknowledgments: HTAi gratefully acknowledges the members of the Coordination Group of the HTAi Community of Practice on AI for HTA for their expertise, feedback, and review of this deliverable: Richard Charter (UK), Rachael Fleurence (USA), Robert Hettle (UK),Pall Jonsson (UK), Sven Klijn (The Netherlands), Bill Malcolm (UK), Rocío Rodríguez López (Spain), Lizzie Shanahan (UK),Dan Singh (Canada), Rebecca Trowman (Australia), Eva Turk (Slovenia), Kristof Vanfraechem (Spain), Siw Waffenschmidt (Germany), Tina Wang (UK).

HTAi Secretariat support:

- Antonio Migliore, Manager of Scientific Initiatives
- Nicola Vicari, Coordinator of Scientific Initiatives

For more information please visit: https://htai.org/engage-with-us/working-groups/htai-copai/

Cite as: Health Technology Assessment international (HTAi). Generative AI Glossary for HTA Professionals. Edmonton, AB: HTAi; 2026.