Advances in prompting techniques

ISPOR Montreal

Sven Klijn

May 14th, 2025

Advances in prompting techniques in the past year

Function calling

Shift from few-shot prompts to contextual examples

Reasoning models

Tool calling

Refinement prompts

Agentic approaches

Self-critique

Extended context windows

RAG

Fine-tuning

Clarification querying

Multimodal prompting

Structured output formatting

Reasoning models

Sep 2024

GPT o1

Reasoning tokens

Chain-of-thought framework

Nov 2024

QwQ-32B-Preview

Jan 2025

DeepSeek R1

Transparent reasoning traces

Low training cost (?)

Feb 2025

Claude 3.7 Sonnet

Hybrid reasoning model

April 2025

Gemini 2.5 Pro

Mixture-of-Experts

April 2025

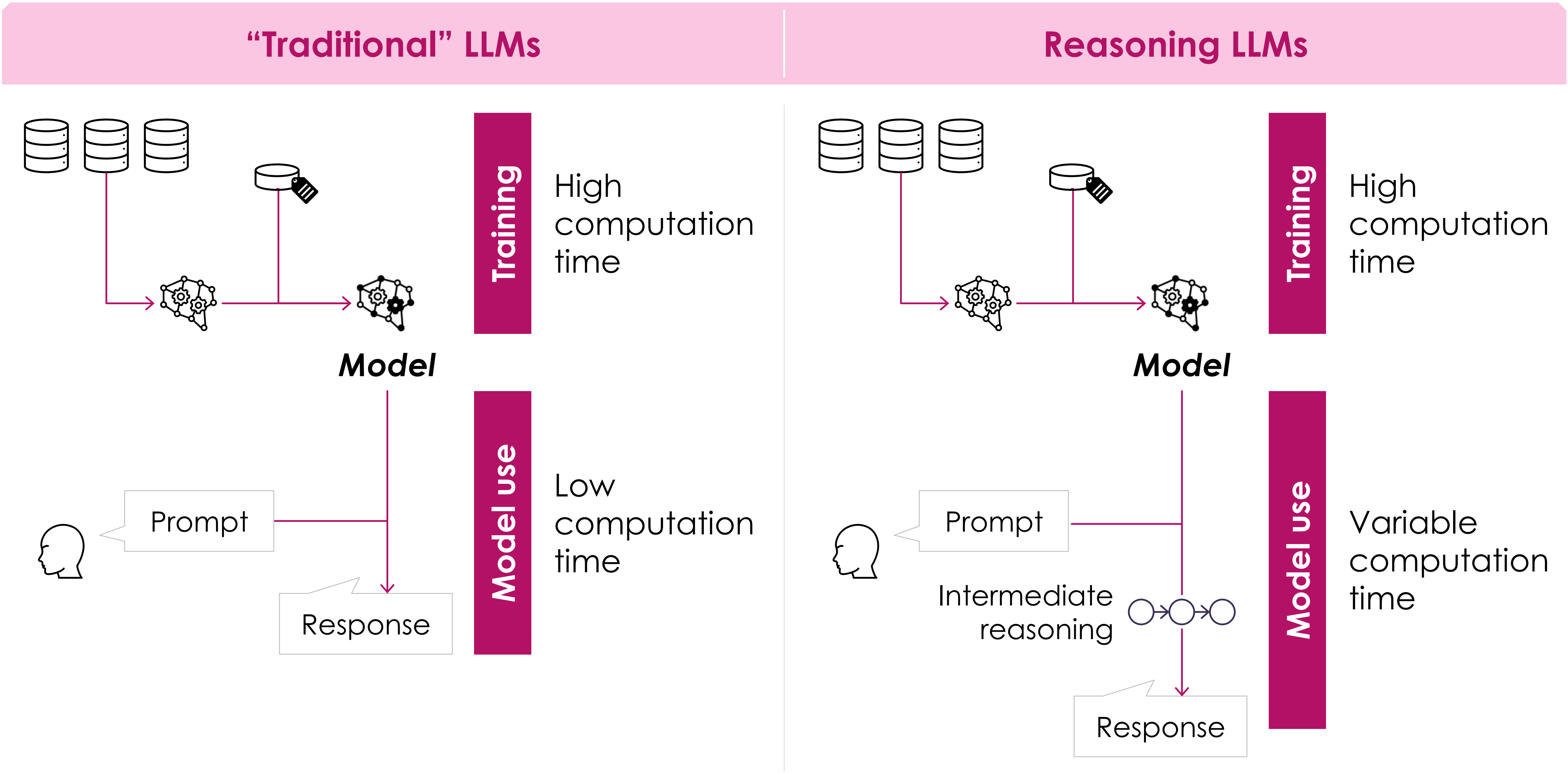
GPT o3

GPT o4-mini

April 2025

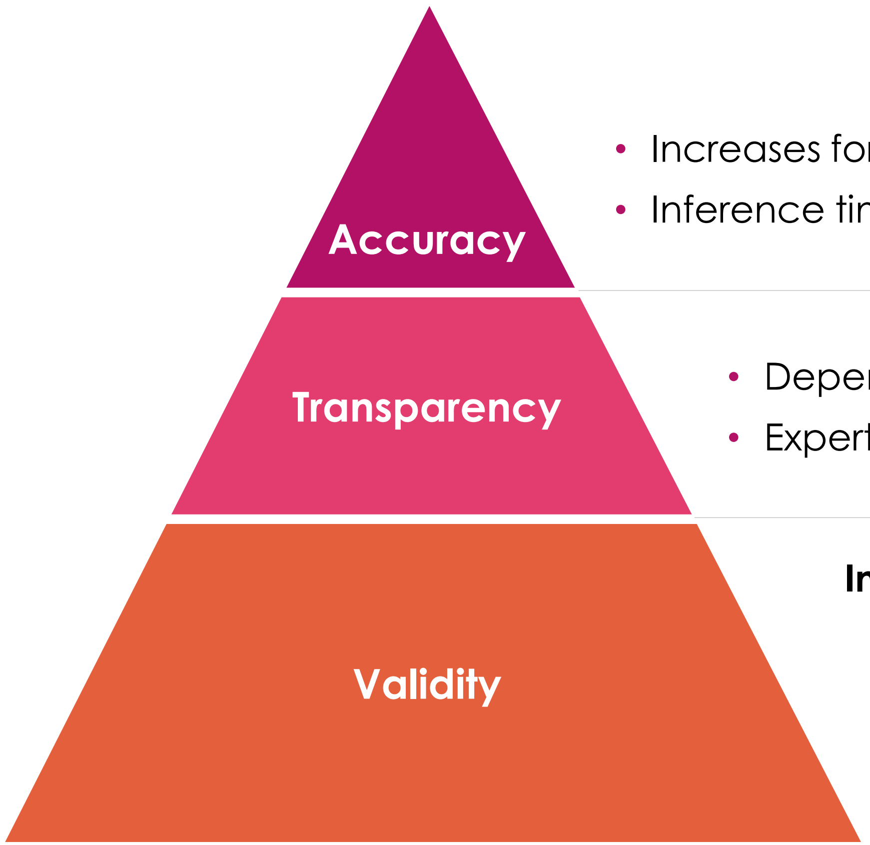
Qwen 3

Hybrid reasoning model



LLM: large language model

Reasoning models: implications for HTA, HEOR and Access



- Increases for multi-step reasoning and calculations
- Inference time computing budget may impact accuracy

- Dependent on visibility of intermediate steps
- Expertise may be required to evaluate logic chains

Internal (repeatability)

- Introduces additional intermediate steps, and potential points of failure
- Logical verification might counteract potential decrease in repeatability

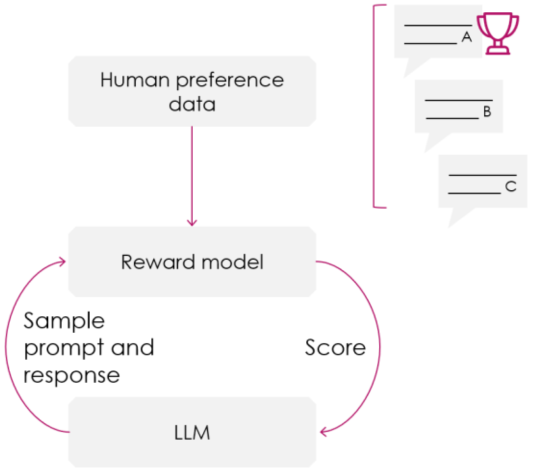
External (generalizability)

- May reduce need for extreme prompt decomposition, increasing generalizability of prompts

Application examples: Multi-step calculations (e.g., price per dose),
Collation of individual assessments (e.g., overall score based on multiple components)

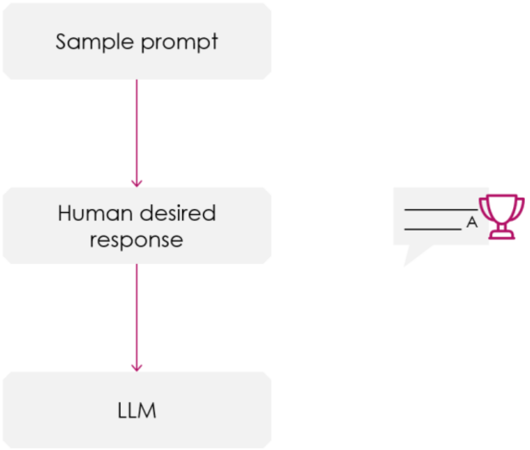
Fine-Tuning

RLHF
Reinforcement Learning from Human Feedback



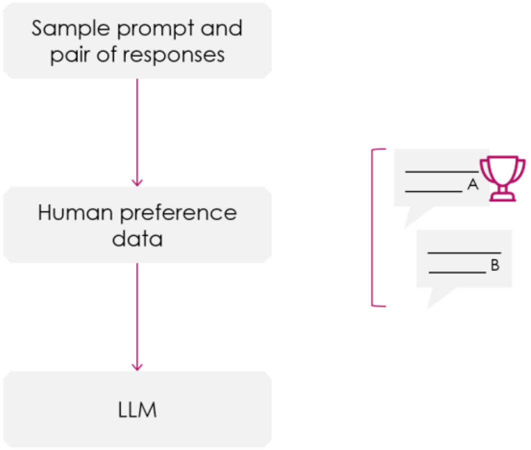
Impossible for closed-source models

SFT
Supervised Fine-Tuning

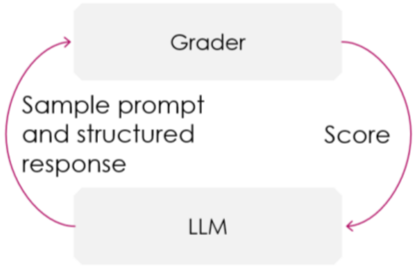


Not feasible due to high data demands

PFT
Preference Fine-Tuning



RFT
Reinforcement Fine-Tuning

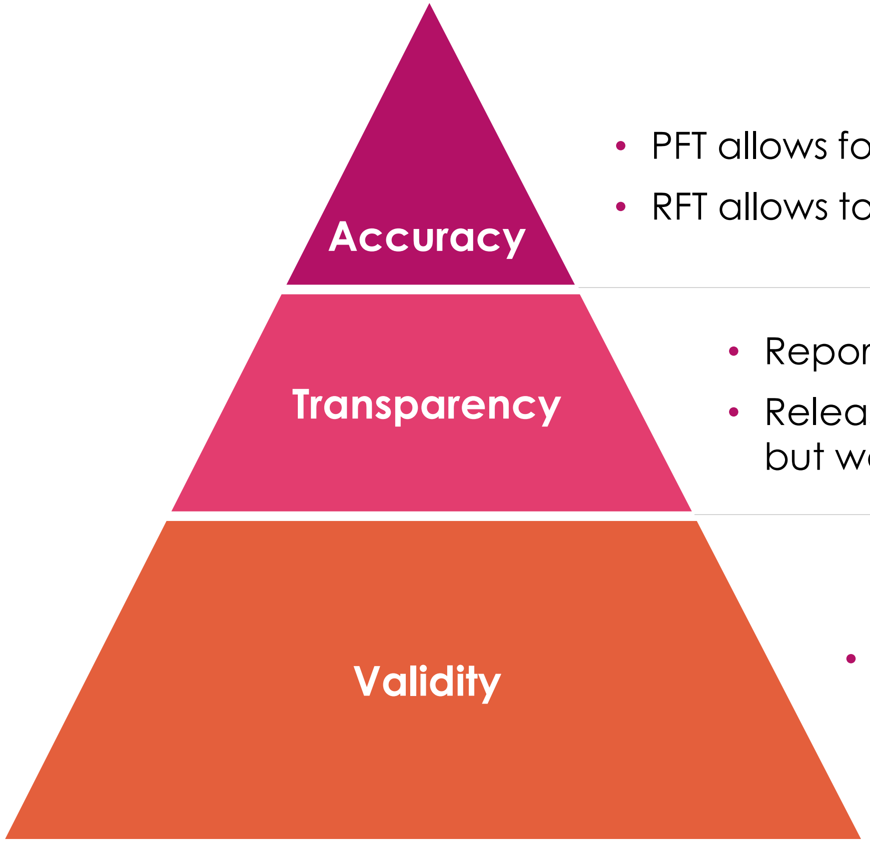


Introduced or made accessible in the past year
Warning: limited evidence available at this point

Comparison of fine-tuning techniques

	RLHF	SFT	PFT	RFT
Objective	Align model with human preferences through reinforcement learning	Improve performance on specific tasks using labeled examples	Align model with human preferences based on human preference labels	Improve performance on specific tasks by reinforcing correct reasoning patterns
Training data format	Reward model fit to human-preference data	Input / output pairs	Input /preferred output / non-preferred output triplets	Input / answer pairs + grader
Amount of training data required	High	High	Moderate	Low / Moderate
Computational resources required	High Iterative reinforcement learning and reward model computation	Moderate Dependent on the size of the dataset	Moderate Less than RLHF but likely higher than SFT	Moderate / High Requires iterative training over the same data points
Use case	When human alignment is critical e.g., content moderation	Single true outcome e.g., custom code format	Subjective outcomes e.g., creative writing	Domains where tasks have objectively correct answers(e.g., HEOR)

Fine-Tuning: implications for HTA, HEOR and Access



Accuracy

- PFT allows for stronger alignment to pre-defined styles
- RFT allows to magnify knowledge on niche domains such as HEOR

Transparency

- Report which training data were used for fine-tuning
- Releasing model weights may look transparent, but weights are almost impossible to interpret by humans

Validity

- Typically increases internal validity (repeatability) at the cost of external validity (generalizability)

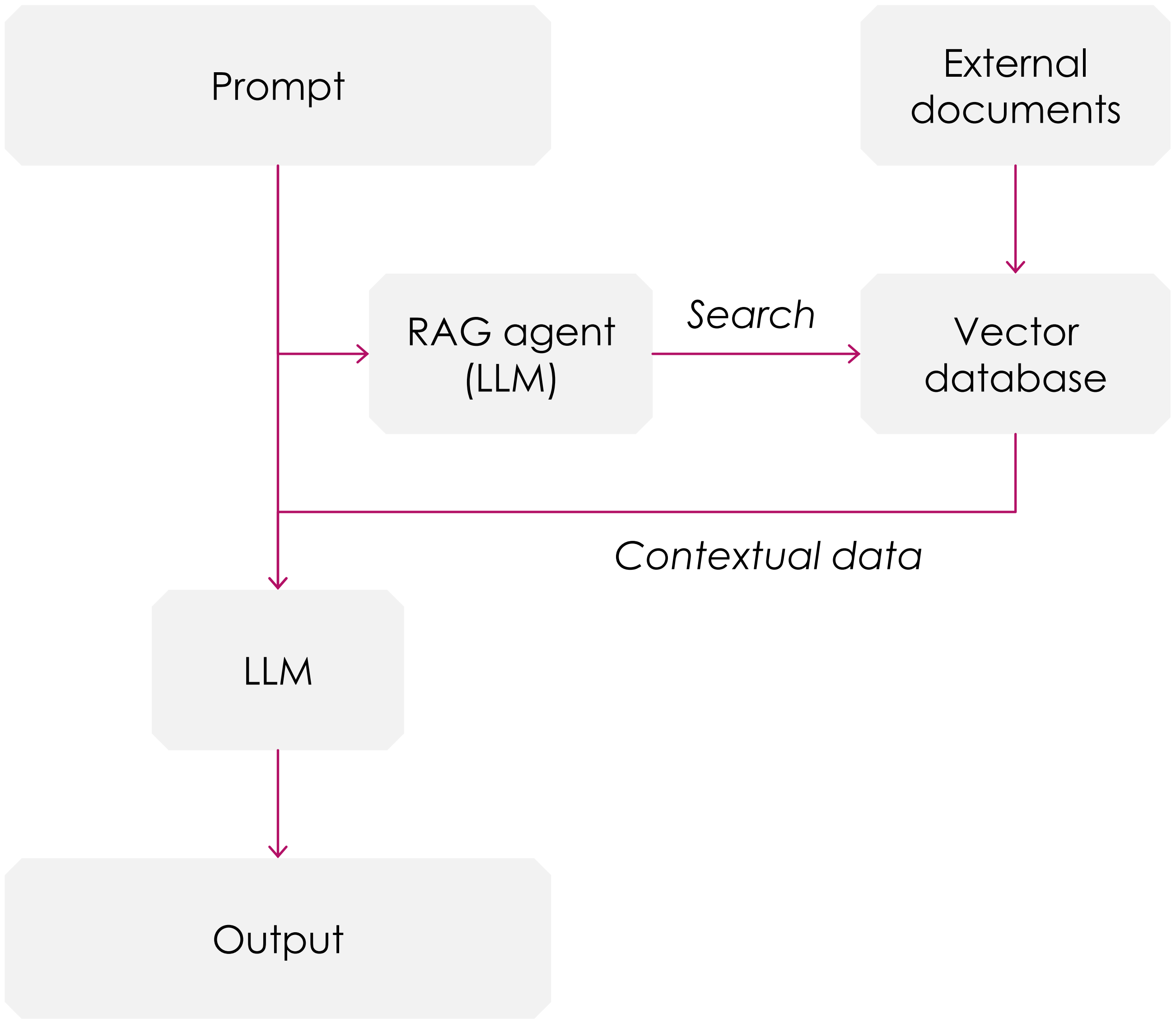
Application examples: Report writing aligned with company style or policies,
Multi-step cost-effectiveness model calculations,
Distillation (train small LLMs to replicate performance of large LLMs, in specific domains)

Retrieval Augmented Generation (RAG)

- RAG provides LLMs with new information that was not included in the model's training data[1]

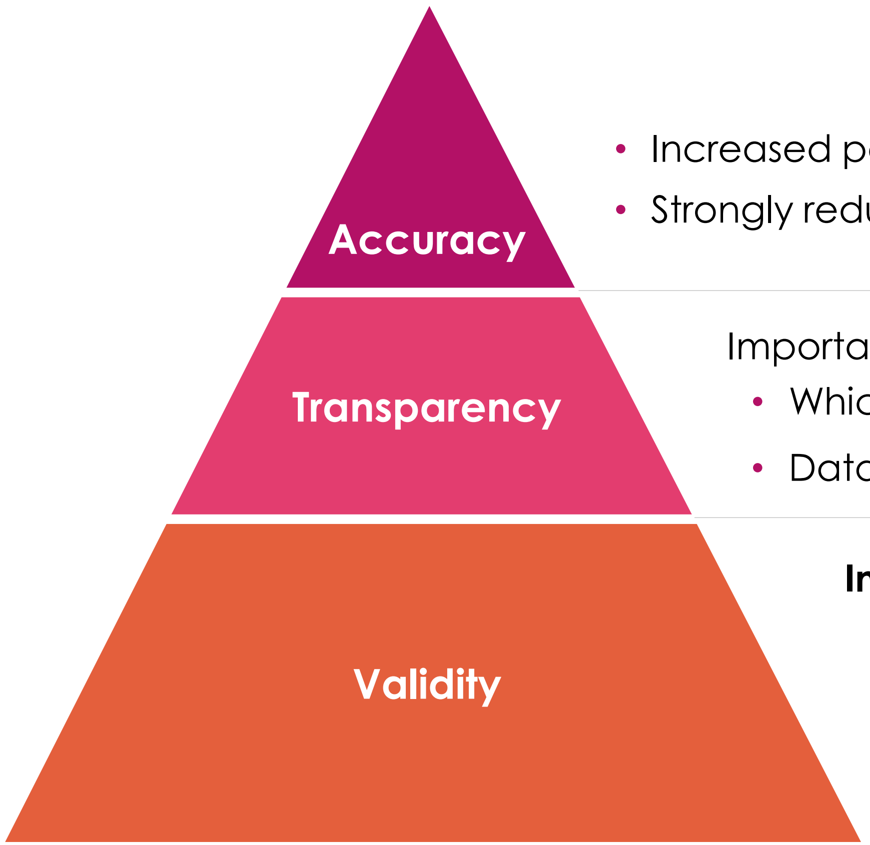
- This is relevant for data that may be:
  - Proprietary
  - Very recent
  - Part of a niche domain

- Data ingestion and retrieval techniques are highly active fields of research



Lewis et al. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. DOI: 10.48550/arXiv.2005.11401

RAG: Implications for HTA, HEOR and Access



- Increased performance for outputs that require niche, recent, and/or proprietary data
- Strongly reduces hallucinations

Important to report:

- Which data were ingested
- Data ingestion and retrieval techniques used

Internal (repeatability)

- Increased, if appropriate data retrieval techniques are used

External (generalizability)

- Depends on data ingested into the database

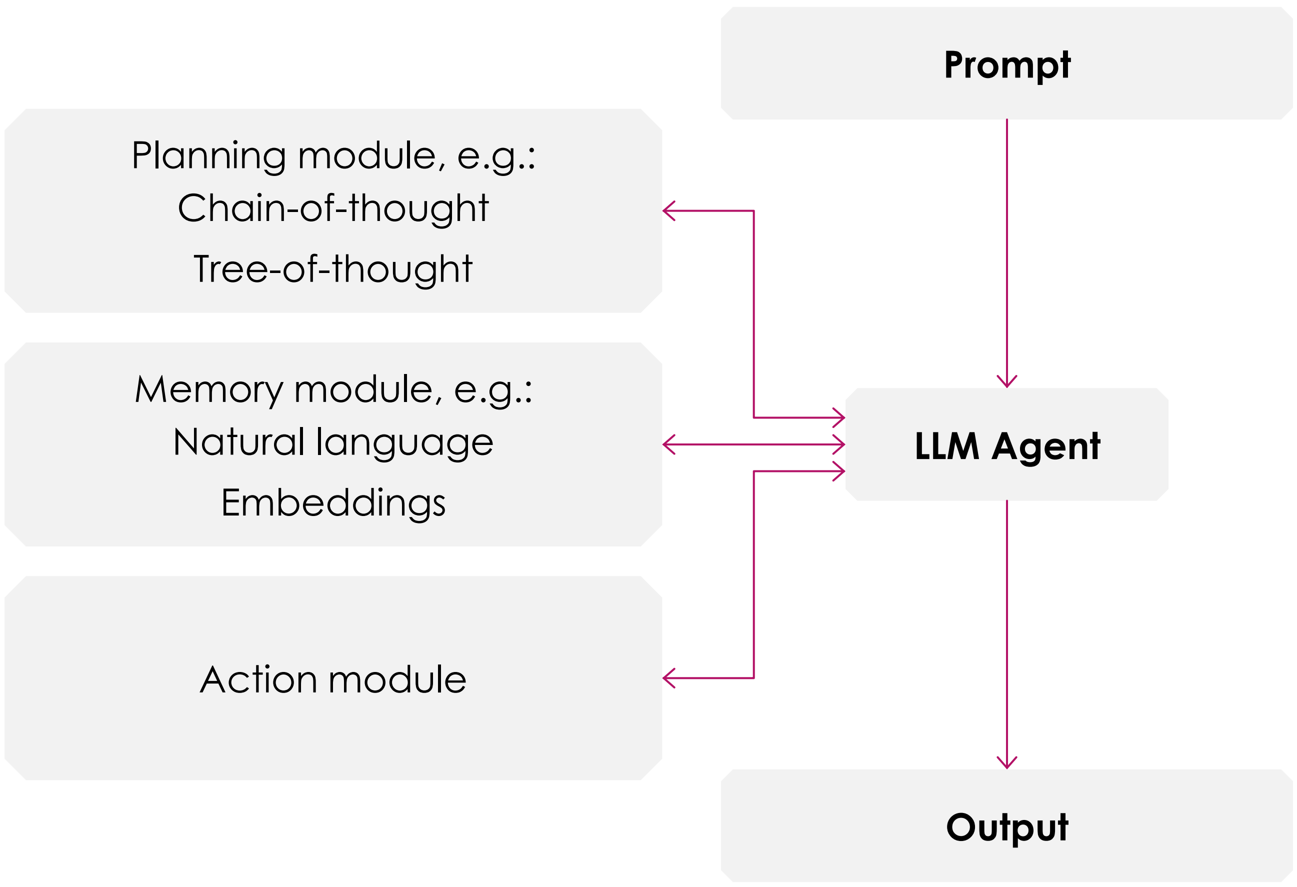
Application examples: Incorporate proprietary knowledge on market dynamics, Reference most recent treatment guidelines

Agentic Approaches: Different levels of agency

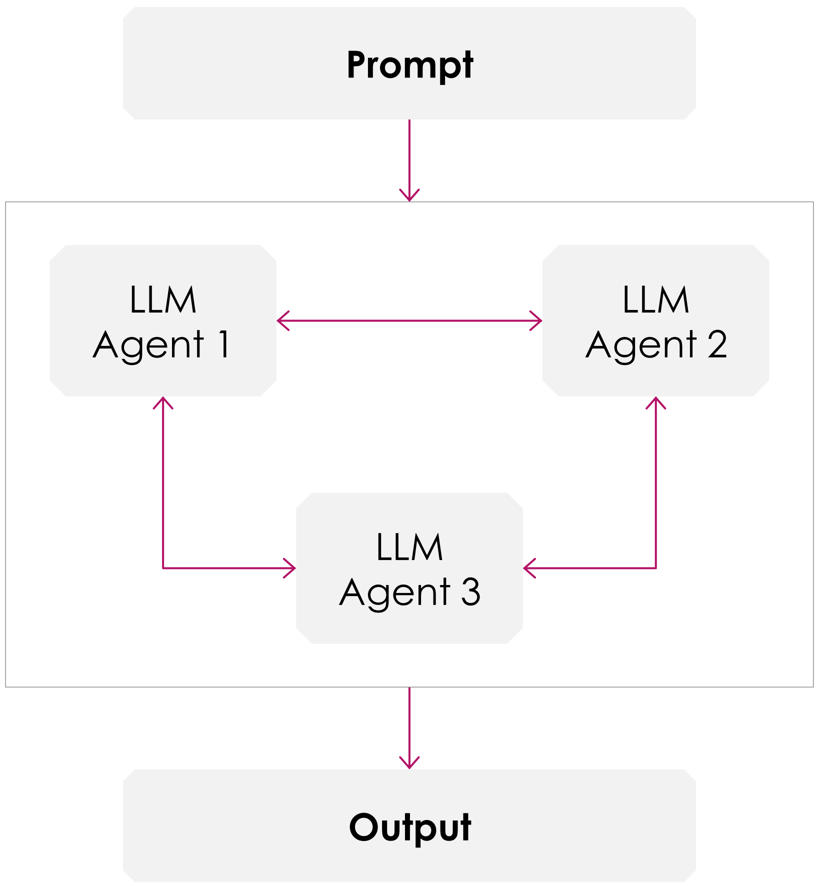
Agency Level	Name	Description	Example Pattern
☆☆☆	Simple Processor	LLM output has no impact on program flow	process_llm_output(llm_response)
★☆☆	Router	LLM output determines an if/else switch	if llm_decision(): • path_a() else: path_b()
★★☆	Tool Caller	LLM output determines function execution	run_function(llm_chosen_tool, llm_chosen_args)
★★★	Multi-step Agent	LLM output controls iteration and program continuation	while llm_should_continue(): execute_next_step()
★★★	Multi-Agent	One agentic workflow can start another agentic workflow	if llm_trigger(): execute_agent()

Source: Roucher (2024), Introduction to agents. Accessed on March 20th, 2025, at https://huggingface.co/docs/smolagents/conceptual_guides/intro_agents

Controller Agent
(Tool Caller or Multi-step Agent)

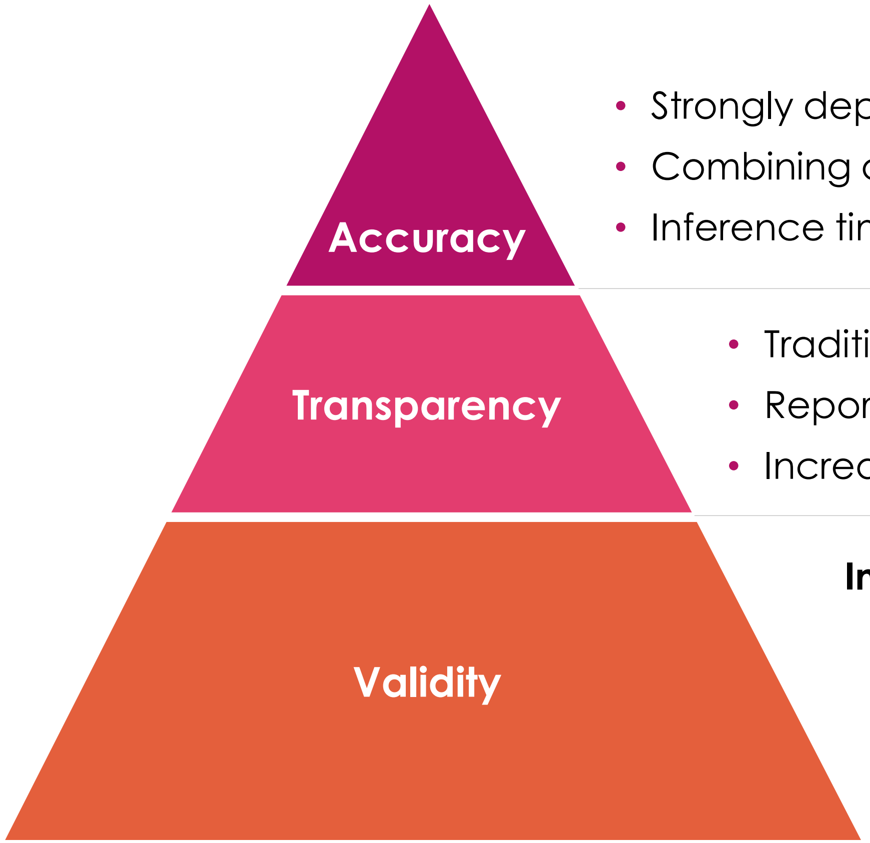


Multi-Agent



Based on: Wang et al. (2024). A Survey on Large Language Model based Autonomous Agents. DOI: 10.48550/arXiv.2308.11432

Agentic Approaches: Implications for HTA, HEOR and Access



- Strongly dependent on the design of the implementation
- Combining agents using different LLMs might counteract model-specific biases
- Inference time computing budget may impact accuracy

- Traditional methods are not sufficient, as e.g., prompts are dynamic
- Reporting trace may be generated as part of the process
- Increased importance of benchmarking for validation

Internal (repeatability)

- Dedicated agents may be designed to increase consistency of outcomes

External (generalizability)

- The basis of an agentic framework may be leveraged for multiple different problems

Application examples: Iterative writing of dossiers,
Simulating review committee discussions

Thank you