CAN GENERATIVE AI RELIABLY IDENTIFY PAYER DECISIONS AND REPLACE MANUAL HTA REVIEW? A COMPARISON OF AI AND HUMAN EXTRACTION OF A FRENCH HTA REPORT

10 NOVEMBER 2026

ISPOR EU 2026 | VIENNA, AUSTRIA

P33

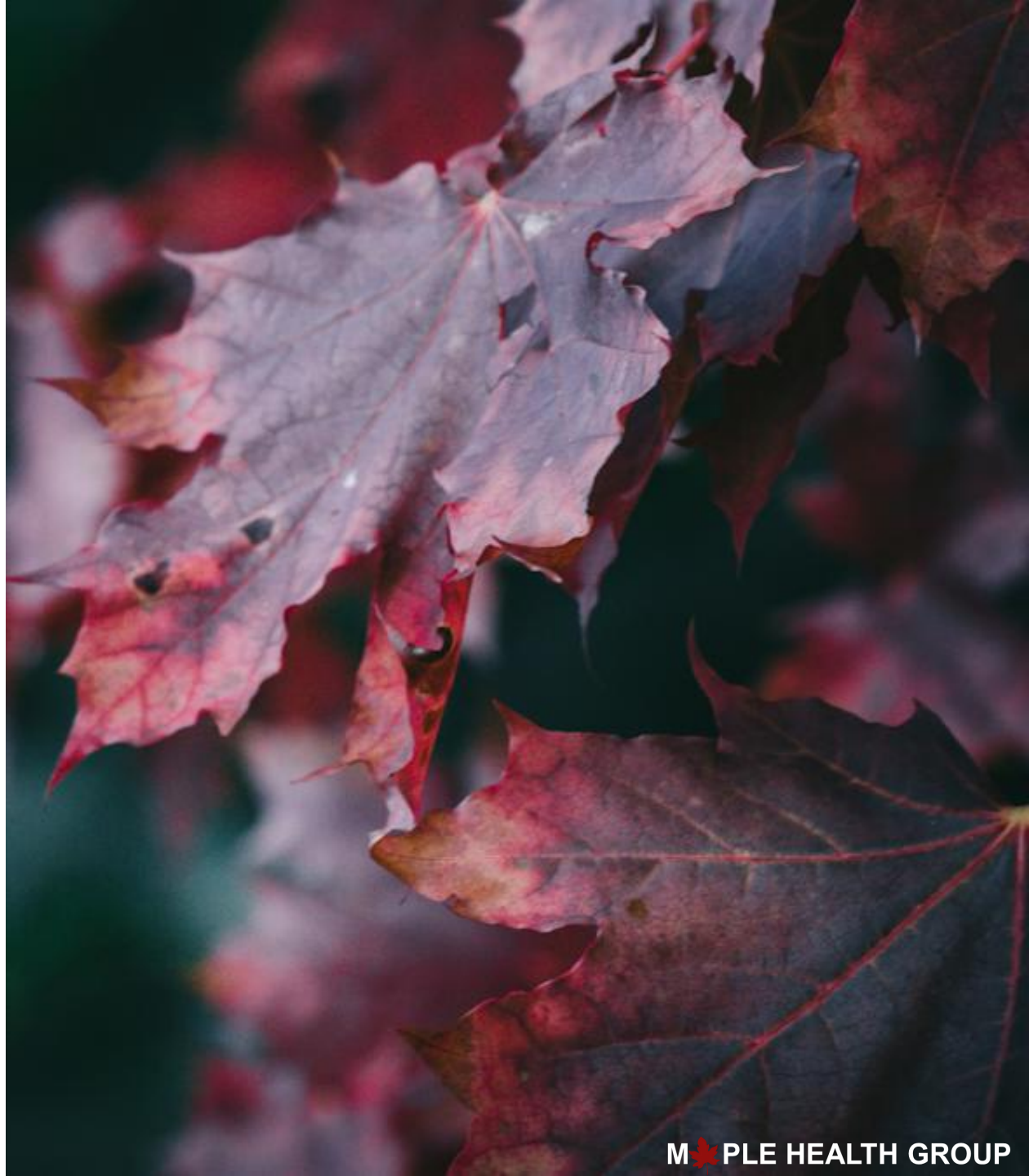
AUTHORS: Rachel Beckerman, Charles Davis, Mirella Dudzic, James Horscroft, Csilla Kinyik-Merena, **Sylwia Lach**, Whitney Longstaff, Janice Stricker-Shaver, Roisin Wynne

sylwia.lach@maplehealthgroup.com

MAPLE HEALTH GROUP

Agenda

1. **RESEARCH OBJECTIVES AND APPROACH**
2. KEY RESEARCH FINDINGS
3. RESEARCH LIMITATIONS
4. Q&A



Why are we here? There is more to HTA than the decision.

HTA reports are a **valuable source of strategic intelligence**, providing insights into payer expectations, decision drivers, and evidence requirements to inform market access and HEOR strategies.

WHY DO HTA FINDINGS MATTER FOR MARKET ACCESS AND HEOR?

- Understand payer expectations and variability across markets
- Anticipate and proactively address expected access barriers
- Assess competitive landscape and implications for your product
- Shape EU JCA, national HTA and P&R strategies
- Guide evidence generation and mitigation to address data gaps
- Support value communication tailored to various payer archetypes

Abbreviations: EU: European Union; HEOR: Health Economics and Outcomes Research; HTA: Health Technology Assessment; JCA: Joint Clinical Assessment.

What if GenAI could turn time-consuming HTA reviews into faster, more efficient analyses?

Extracting and interpreting HTA findings can be time-consuming and resource-intensive. GenAI offers the **potential to accelerate, standardise, and scale evidence extraction and synthesis**, however, maintaining accuracy and reliability remains critical for well-informed decisions.

WHAT ARE THE CURRENT CHALLENGES AND OPPORTUNITIES FOR GENERATIVE AI?

SELECTED CHALLENGES OF MANUAL HTA REVIEW

- High volume of HTA documents across markets
- Time-intensive reviews often under rapid turnaround
- Variability in HTA reporting across countries, agencies, and therapeutic areas
- Potentially repetitive effort across different projects

SELECTED GENERATIVE AI OPPORTUNITIES

- Accelerate HTA report extraction for increased time efficiency
- Standardise analyses across countries and agencies
- Support evidence synthesis to identify cross-market patterns and strategic insights
- Improve consistency and traceability of findings

Abbreviations: GenAI: Generative Artificial Intelligence; HEOR: Health Economics and Outcomes Research; HTA: Health Technology Assessment.

Taking GenAI further: From identifying HTA analogues to understanding the decisions and rationale behind them.

Building on Part 1, this research shifts the focus from identifying HTA analogues to exploring GenAI's **accuracy and reliability in extracting, interpreting, and structuring information** from HTA reports.

IDENTIFICATION	*INTERPRETATION*	*SYNTHESIS*

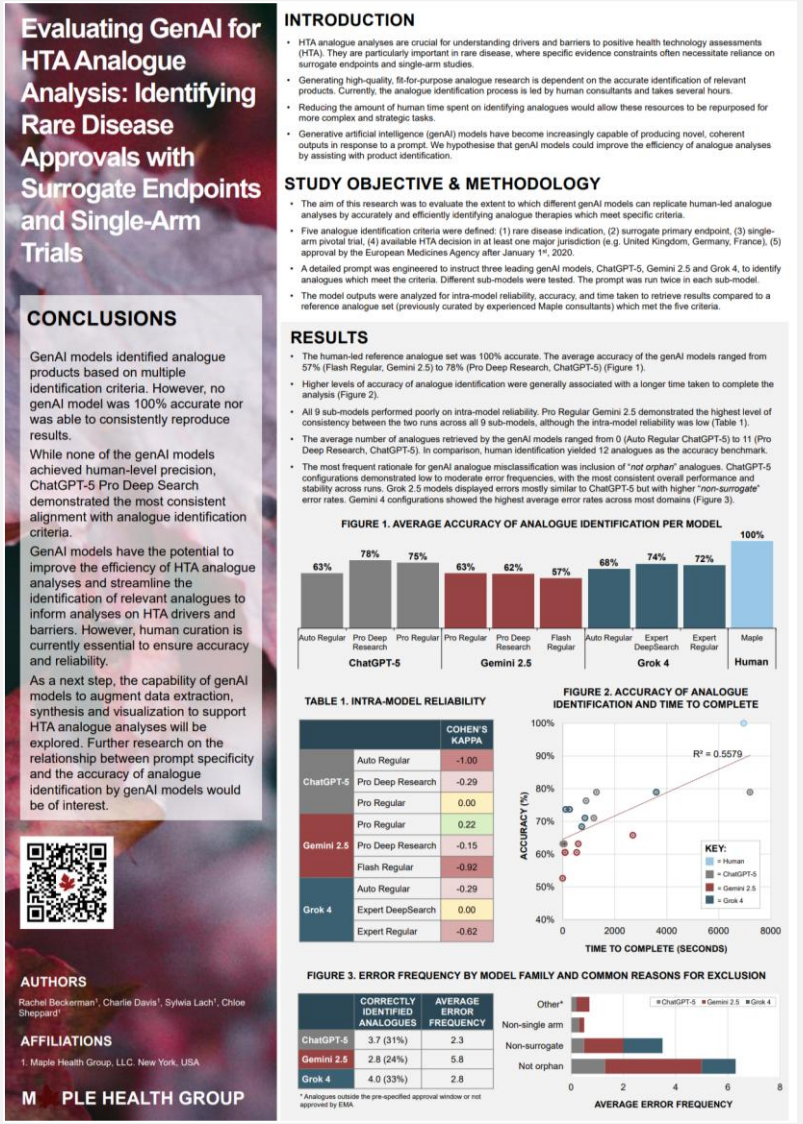
PART 1
Can GenAI models identify relevant HTA analogues?

PART 2
Can GenAI identify and explain the rationale behind HTA decisions?

CURRENT STEP

PART 3
Can GenAI analyse and synthesise HTA evidence into actionable insights?

PART 1 OF THE RESEARCH



Evaluating GenAI for HTA Analogue Analysis: Identifying Rare Disease Approvals with Surrogate Endpoints and Single-Arm Trials

CONCLUSIONS

GenAI models identified analogue products based on multiple identification criteria. However, no genAI model was 100% accurate nor was able to consistently reproduce results.

While none of the genAI models achieved human-level precision, ChatGPT-5 Pro Deep Search demonstrated the most consistent alignment with analogue identification criteria.

GenAI models have the potential to improve the efficiency of HTA analogue analyses and streamline the identification of relevant analogues to inform analyses on HTA drivers and barriers. However, human curation is currently essential to ensure accuracy and reliability.

As a next step, the capability of genAI models to augment data extraction, synthesis and visualization to support HTA analogue analyses will be explored. Further research on the relationship between prompt specificity and the accuracy of analogue identification by genAI models would be of interest.

AUTHORS

Rachel Beckerman[1], Charlie Davis[1], Sylwia Lach[1], Chloe Sheppard[1]

AFFILIATIONS

1. Maple Health Group, LLC. New York, USA

M PLE HEALTH GROUP

INTRODUCTION

- HTA analogue analyses are crucial for understanding drivers and barriers to positive health technology assessments (HTA). They are particularly important in rare disease, where specific evidence constraints often necessitate reliance on surrogate endpoints and single-arm studies.
- Generating high-quality, fit-for-purpose analogue research is dependent on the accurate identification of relevant products. Currently, the analogue identification process is led by human consultants and takes several hours.
- Reducing the amount of human time spent on identifying analogues would allow these resources to be repurposed for more complex and strategic tasks.
- Generative artificial intelligence (genAI) models have become increasingly capable of producing novel, coherent outputs in response to a prompt. We hypothesise that genAI models could improve the efficiency of analogue analyses by assisting with product identification.

STUDY OBJECTIVE & METHODOLOGY

- The aim of this research was to evaluate the extent to which different genAI models can replicate human-led analogue analyses by accurately and efficiently identifying analogue therapies which meet specific criteria.
- Five analogue identification criteria were defined: (1) rare disease indication, (2) surrogate primary endpoint, (3) single-arm pivotal trial, (4) available HTA decision in at least one major jurisdiction (e.g. United Kingdom, Germany, France), (5) approval by the European Medicines Agency after January 1st, 2020.
- A detailed prompt was engineered to instruct three leading genAI models, ChatGPT-5, Gemini 2.5 and Grok 4, to identify analogues which meet the criteria. Different sub-models were tested. The prompt was run twice in each sub-model.
- The model outputs were analyzed for intra-model reliability, accuracy, and time taken to retrieve results compared to a reference analogue set (previously curated by experienced Maple consultants) which met the five criteria.

RESULTS

- The human-led reference analogue set was 100% accurate. The average accuracy of the genAI models ranged from 57% (Flash Regular, Gemini 2.5) to 78% (Pro Deep Research, ChatGPT-5) (Figure 1).
- Higher levels of accuracy of analogue identification were generally associated with a longer time taken to complete the analysis (Figure 2).
- All 9 sub-models performed poorly on intra-model reliability. Pro Regular Gemini 2.5 demonstrated the highest level of consistency between the two runs across all 9 sub-models, although the intra-model reliability was low (Table 1).
- The average number of analogues retrieved by the genAI models ranged from 0 (Auto Regular ChatGPT-5) to 11 (Pro Deep Research, ChatGPT-5). In comparison, human identification yielded 12 analogues as the accuracy benchmark.
- The most frequent rationale for genAI analogue misclassification was inclusion of "*not orphan*" analogues. ChatGPT-5 configurations demonstrated low to moderate error frequencies, with the most consistent overall performance and stability across runs. Grok 2.5 models displayed errors mostly similar to ChatGPT-5 but with higher "*non-surrogate*" error rates. Gemini 4 configurations showed the highest average error rates across most domains (Figure 3).

FIGURE 1. AVERAGE ACCURACY OF ANALOGUE IDENTIFICATION PER MODEL

TABLE 1. INTRA-MODEL RELIABILITY

		COHEN'S KAPPA
ChatGPT-5	Auto Regular	-1.00
	Pro Deep Research	-0.29
	Pro Regular	0.00
Gemini 2.5	Pro Regular	0.22
	Pro Deep Research	-0.15
	Flash Regular	-0.92
Grok 4	Auto Regular	-0.29
	Expert DeepSearch	0.00
	Expert Regular	-0.62

FIGURE 2. ACCURACY OF ANALOGUE IDENTIFICATION AND TIME TO COMPLETE

FIGURE 3. ERROR FREQUENCY BY MODEL FAMILY AND COMMON REASONS FOR EXCLUSION

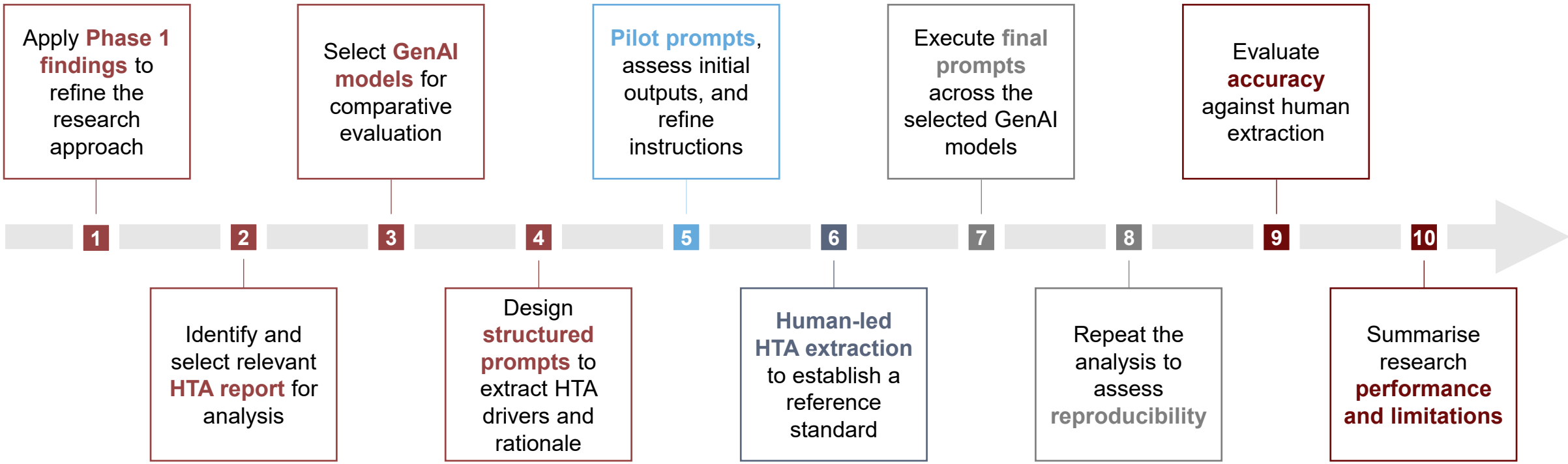
	CORRECTLY IDENTIFIED ANALOGUES	AVERAGE ERROR FREQUENCY
ChatGPT-5	3.7 (31%)	2.3
Gemini 2.5	2.8 (24%)	5.8
Grok 4	4.0 (33%)	2.8

* Analogues outside the pre-specified approval window or not approved by EMA

* Research presented at ISPOR EU 2025 concluded that, while GenAI models could identify analogue products, none achieved 100% accuracy, highlighting the continued need for human validation. Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

Putting GenAI to the test: A structured approach to assessing accuracy, reliability, and reproducibility.

GenAI offers the potential to transform time-consuming HTA reviews into faster analyses, but can it deliver and **reliably identify key decision drivers and their underlying rationale**?



Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

Inside the GenAI evaluation: GenAI models, HTA case study, and assessment setup.

Three GenAI models were evaluated using a **standardised prompt and two independent runs per model, with outputs benchmarked against human extraction** to assess accuracy and reproducibility. The outputs were generated in an Excel format to facilitate comparison across models, runs and against the human extraction.

VALOCTOCOGENE ROXAPARVOVEC IN SEVERE HAEMOPHILIA A

HTA body: HAS | **Report date**: 2023 | **Version used**: French

THREE GENERATIVE AI MODELS EVALUATED

ChatGPT 5.5 Pro

Gemini 3.1 Pro Extended

Claude Opus 4.8 Max

STANDARDISED PROMPT
Detailed extraction instructions, exclusions, and output structure

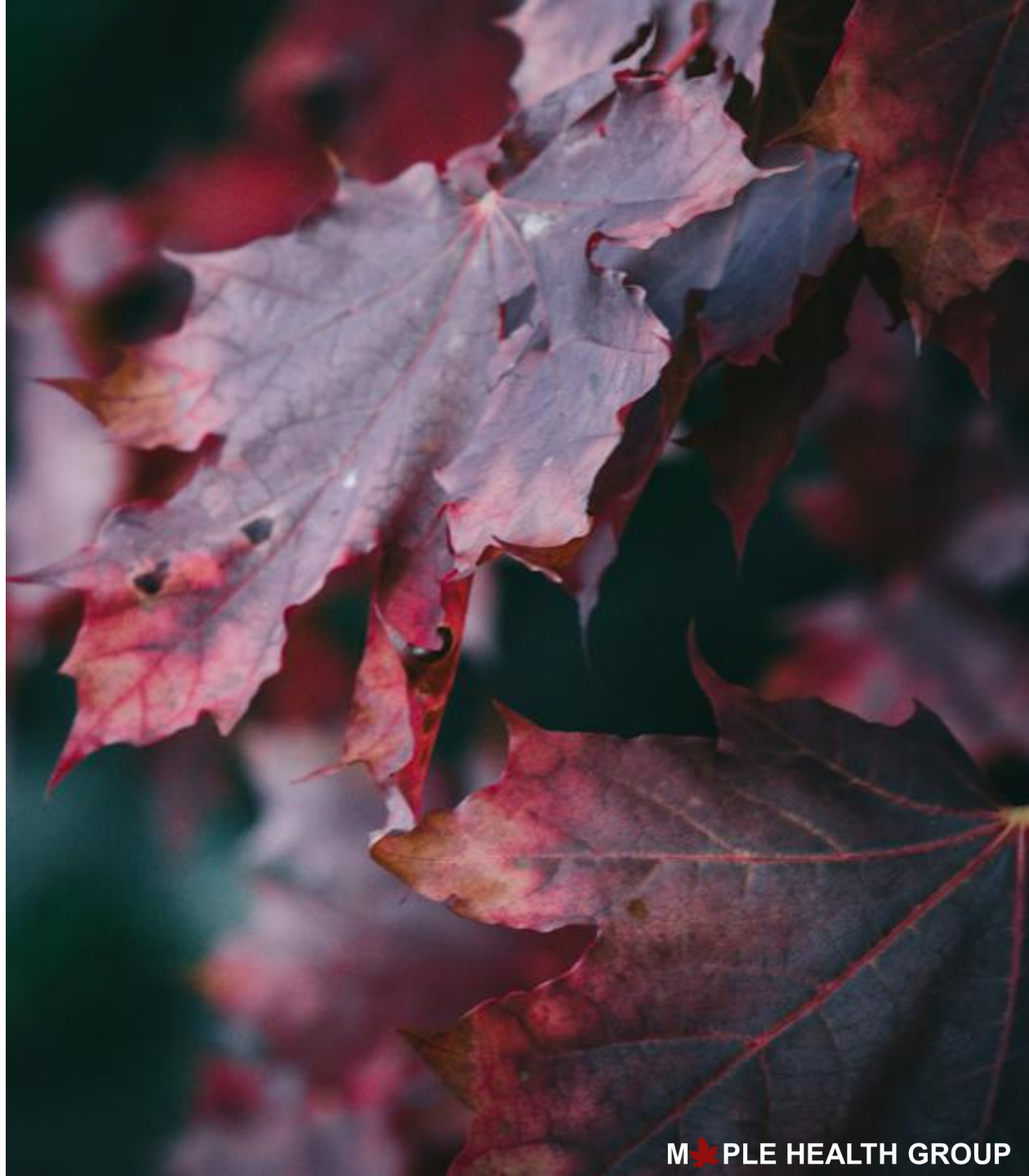
TWO RUNS PER MODEL
Six outputs generated to assess reproducibility

HUMAN REFERENCE
GenAI outputs compared against human-led extraction

NOTE: Prompts were run on June 9th, 2026, with one-hour difference between runs. The original HAS PDF uploaded to each GenAI model. Abbreviations: GenAI: Generative Artificial Intelligence; HAS: Haute Autorité de Santé (High Authority for Health, France); HTA: Health Technology Assessment.

Agenda

1. RESEARCH OBJECTIVES AND APPROACH
2. **KEY RESEARCH FINDINGS**
3. RESEARCH LIMITATIONS
4. Q&A



GenAI extractions were evaluated across pre-defined assessment domains and questions to assess the accuracy and completeness of extracted findings.

ASSESSMENT DOMAINS AND RESEARCH QUESTIONS

HTA RECOMMENDATION AND RATIONALE

DISEASE BURDEN AND UNMET NEEDS

SUBMITTED EVIDENCE BASE

DID THE MODEL CORRECTLY CAPTURE:
The exact HTA outcome and rationale? | Disease burden and unmet needs? | Submitted trial(s), ITC(s) and RWE?

POPULATION

INTERVENTION

COMPARATOR

OUTCOMES

STUDY DESIGN

DID THE MODEL CORRECTLY CAPTURE:
HTA negative commentary? | HTA positive commentary? | Only explicit HTA insights with no own interpretation?

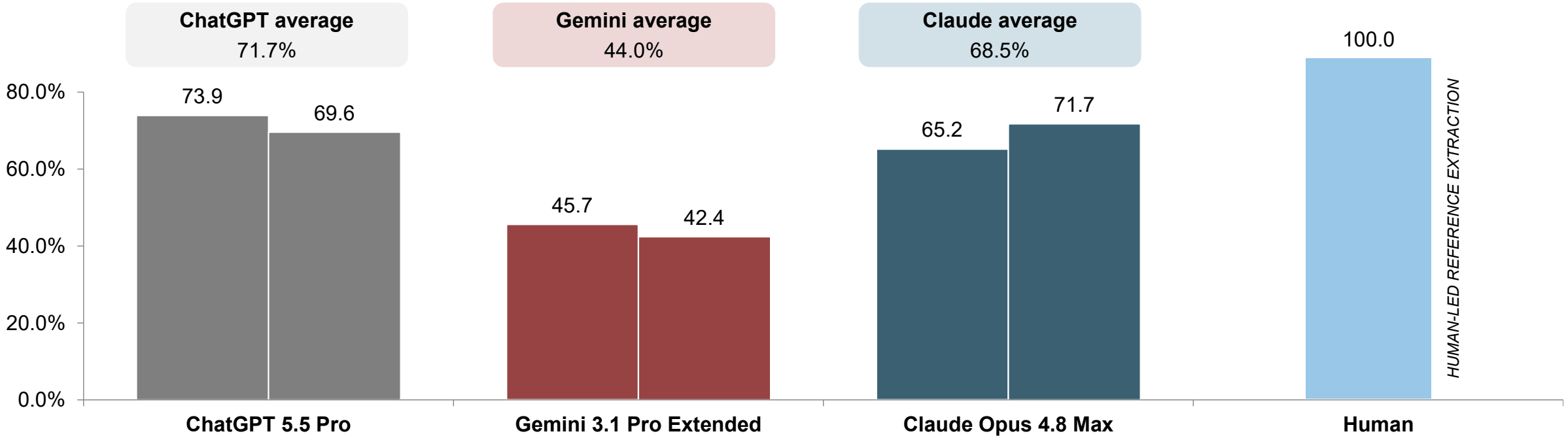
Each criterion was assessed against the human reference extraction qualitatively and quantitatively using a 0-2 scoring scale (0 = incorrect, 1 = partially correct, 2 = fully correct). Quantitative scores were averaged across the two GenAI runs and normalised to a 0-100% scale.

Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

GenAI often missed important details and nuances from the HTA reports, with accuracy varying across models, highlighting the need for a subsequent validation.

Across all three models, GenAI extractions were **partially correct on average, generally capturing the direction of HTA findings but missing details or nuances compared with human extraction**. Results were relatively consistent across runs, although highlighted the need for human validation to ensure accuracy and completeness.

FIGURE 1. NORMALISED AVERAGE ACCURACY SCORE (%) OF GENERATIVE AI AND HUMAN EXTRACTION



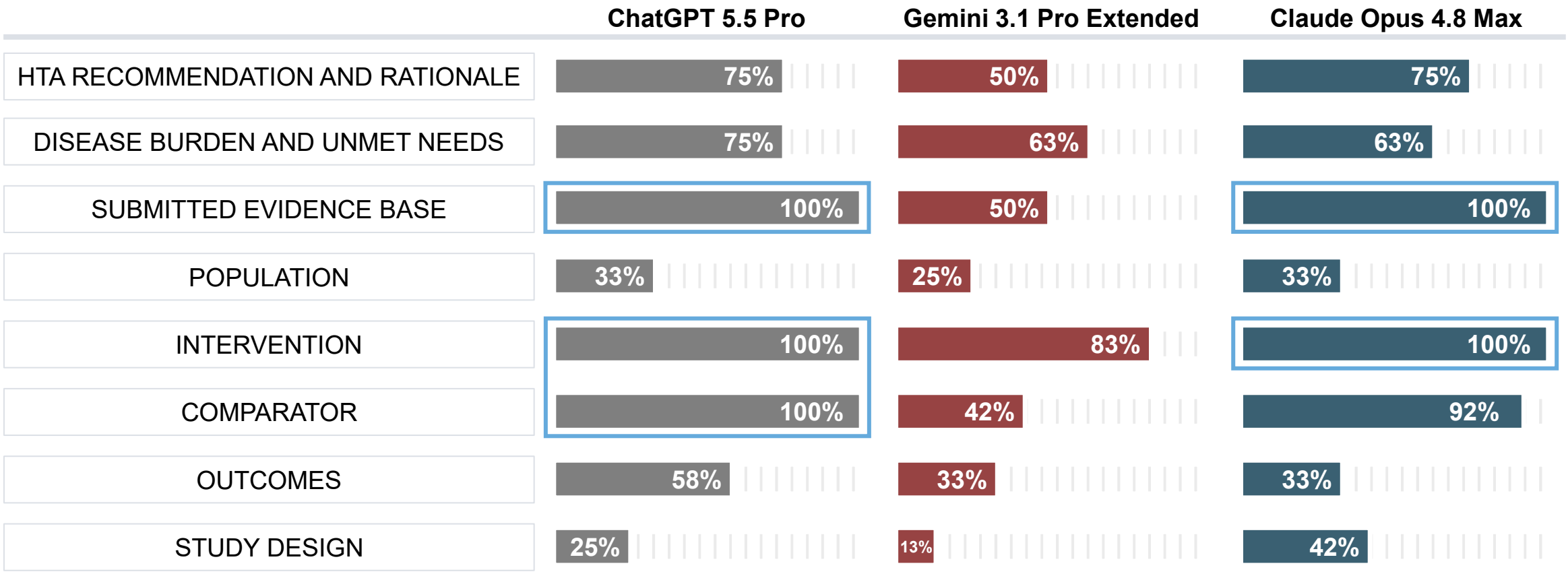
Extraction accuracy was calculated as a percentage of the maximum achievable score across applicable assessment criteria, using a two-point scale: 0 = incorrect (0%), 1 = partially correct (50%), and 2 = fully correct (100%).

NOTE: Accuracy summary was based on a measurement of GenAI performance, ranging from 0 (incorrect information) to 2 (correct information) and was assessed twice per each prompt run per GenAI model. Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

GenAI extraction accuracy varied across models and assessment domains, with challenges in capturing nuanced HTA findings and a sufficient decision rationale.

GenAI extraction **accuracy varied considerably across models and assessment domains**, with ChatGPT generally achieving the highest scores. All models showed most limitations in capturing commentary on the payer perception on the population, outcomes and study design, which remain critical insights for any HTA analysis.

FIGURE 2. NORMALISED AVERAGE ACCURACY SCORE (%) OF GENERATIVE AI EXTRACTION PER ASSESSMENT DOMAIN



NOTE: Accuracy summary was based on a measurement of GenAI performance, ranging from 0 (incorrect information) to 2 (correct information) and was assessed twice per each prompt run per GenAI model, an average rating from both prompt runs was applied. Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

1 CORE HTA FINDINGS WERE GENERALLY CAPTURED

GenAI identified main HTA conclusions, but completeness of supporting information varied across models

HTA DECISIONS

- ✓ All models captured reimbursement recommendation and SMR/ASMR ratings
- ✗ Claude and Gemini omitted reassessment or additional data requests; clinical uncertainties were incomplete

DISEASE BURDEN & UNMET NEEDS

- ✓ All models captured key disease manifestations and clinical burden
- ✗ Unmet need considerations were less complete; Gemini overstated the effectiveness of existing treatments

SUBMITTED EVIDENCE BASE

- ✓ ChatGPT and Claude identified the submitted studies and analyses
- ✗ Gemini omitted the PK model, cited an incorrect comparator study, and incompletely captured an ITC

INTERVENTION

- ✓ All models identified the main treatment opportunities and limitations
- ✗ Gemini omitted enhanced monitoring requirements described in the product information

IMPLICATIONS FOR HTA RESEARCH

→ GenAI may support initial HTA extraction, but human validation remains essential to ensure complete and accurate reporting

Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment; ITC: Indirect Treatment Comparison; PK: Pharmacokinetics; SMR/ASMR: Medical Service Rendered / Improvement of Medical Service Rendered.

2 HTA COMMENTARY WAS NOT ALWAYS CORRECTLY INTERPRETED

GenAI may misrepresent the HTA's actual judgement, resulting in inaccurate interpretations of HTA conclusions

NEUTRAL INFORMATION PRESENTED AS POSITIVE

Trial population alignment with the indication, sample size or follow-up considered as positive commentary rather than statement

REFERENCE:

Study characteristics were reported descriptively, while HAS focused on population heterogeneity and uncertain generalisability critique

POSITIVE COMMENTARY MISSED

Gemini overlooked HAS's positive assessment of comparator relevance, incorrectly reporting its absence

REFERENCE:

FVIII concentrates and emicizumab were explicitly identified as clinically relevant comparators for assessing product's added therapeutic benefit

HTA CONCLUSIONS MISINTERPRETED

Gemini misinterpreted ASMR V as evidence that the treatment provides no clinical benefit over existing therapies

REFERENCE:

Added therapeutic benefit was not demonstrated, given the lack of robust comparative evidence, trial limitations and uncertainties

IMPLICATIONS FOR HTA RESEARCH

GenAI outputs should clearly distinguish explicit HTA judgements from factual evidence descriptions and model-generated interpretations to avoid misrepresenting payer perceptions and conclusions

Abbreviations: ASMR: Amélioration du Service Médical Rendu; FVIII: Factor VIII; GenAI: Generative Artificial Intelligence; HAS: Haute Autorité de Santé (High Authority for Health, France); HTA: Health Technology Assessment.

3

METHODOLOGICAL LIMITATIONS WERE INCOMPLETELY ASSESSED

GenAI identified key trial limitations but inconsistently reported methodological concerns of indirect comparisons

PIVOTAL CLINICAL TRIAL

- ✓ All models identified the non-comparative, open-label design and associated risk of bias
- ✗ Endpoint subjectivity and uncertainty in interpreting efficacy estimates, were incompletely captured

INDIRECT COMPARISONS

- ✓ Some models captured positive commentary on MAIC quality, one identified methodological limitations
- ✗ Most methodological limitations were missed, including those of the second indirect comparison by all models

ChatGPT 5.5 Pro

- ✓ Key pivotal trial limitations captured
- ✗ Lack of any ITC methodological critique

Gemini 3.1 Pro Extended

- ✓ Selected pivotal trial limitations captured
- ✗ Missed all ITC methodological considerations

Claude Opus 4.8 Max

- ✓ Most detailed methodological critique
- ✗ Incomplete methodological critique, with one ITC omitted

IMPLICATIONS FOR HTA RESEARCH

→ Targeted methodological review is needed to adequately characterise the limitations and uncertainties of comparative evidence

Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment; ITC: Indirect Treatment Comparison; MAIC: Matching-Adjusted Indirect Comparison.

4

MODEL PERFORMANCE VARIED ACROSS ASSESSMENT DOMAINS

ChatGPT and Claude achieved similar overall accuracy, but performance varied across assessment domains

ChatGPT 5.5 Pro

71.7%

OVERALL ACCURACY

- ✓ Highest or joint-highest accuracy in 7 of 8 domains, particularly submitted evidence, intervention and comparators
- ✗ Incomplete decision rationale and methodological limitations of indirect comparisons

Gemini 3.1 Pro Extended

44.0%

OVERALL ACCURACY

- ✓ Captured high-level HTA conclusions and main findings on intervention and disease burden characteristics
- ✗ Lowest accuracy across most domains, greater omissions, misinterpretations or inaccuracies

Claude Opus 4.8 Max

68.5%

OVERALL ACCURACY

- ✓ Strong evidence identification and highest study design accuracy, with most detailed methodological commentary
- ✗ Methodological critique remained incomplete and, at times, unnecessarily detailed

IMPLICATIONS FOR HTA RESEARCH

→ Overall accuracy alone may mask domain-specific weaknesses; model selection should consider performance across individual HTA assessment domains and consistency between runs, with broader testing required

Abbreviations: HTA: Health Technology Assessment.

GenAI can support HTA extraction, but human expertise remains essential for interpreting the evidence and the rationale behind payer decisions.

KEY RESEARCH TAKEAWAYS

- **CORE HTA FINDINGS WERE GENERALLY CAPTURED**
- **HTA COMMENTARY WAS NOT ALWAYS CORRECTLY INTERPRETED**
- **METHODOLOGICAL LIMITATIONS WERE INCOMPLETELY ASSESSED**
- **MODEL PERFORMANCE VARIED ACROSS ASSESSMENT DOMAINS**

FUTURE RESEARCH PRIORITIES

REFINE PROMPTS
Distinguish explicit HTA judgements from factual descriptions and request supporting rationale

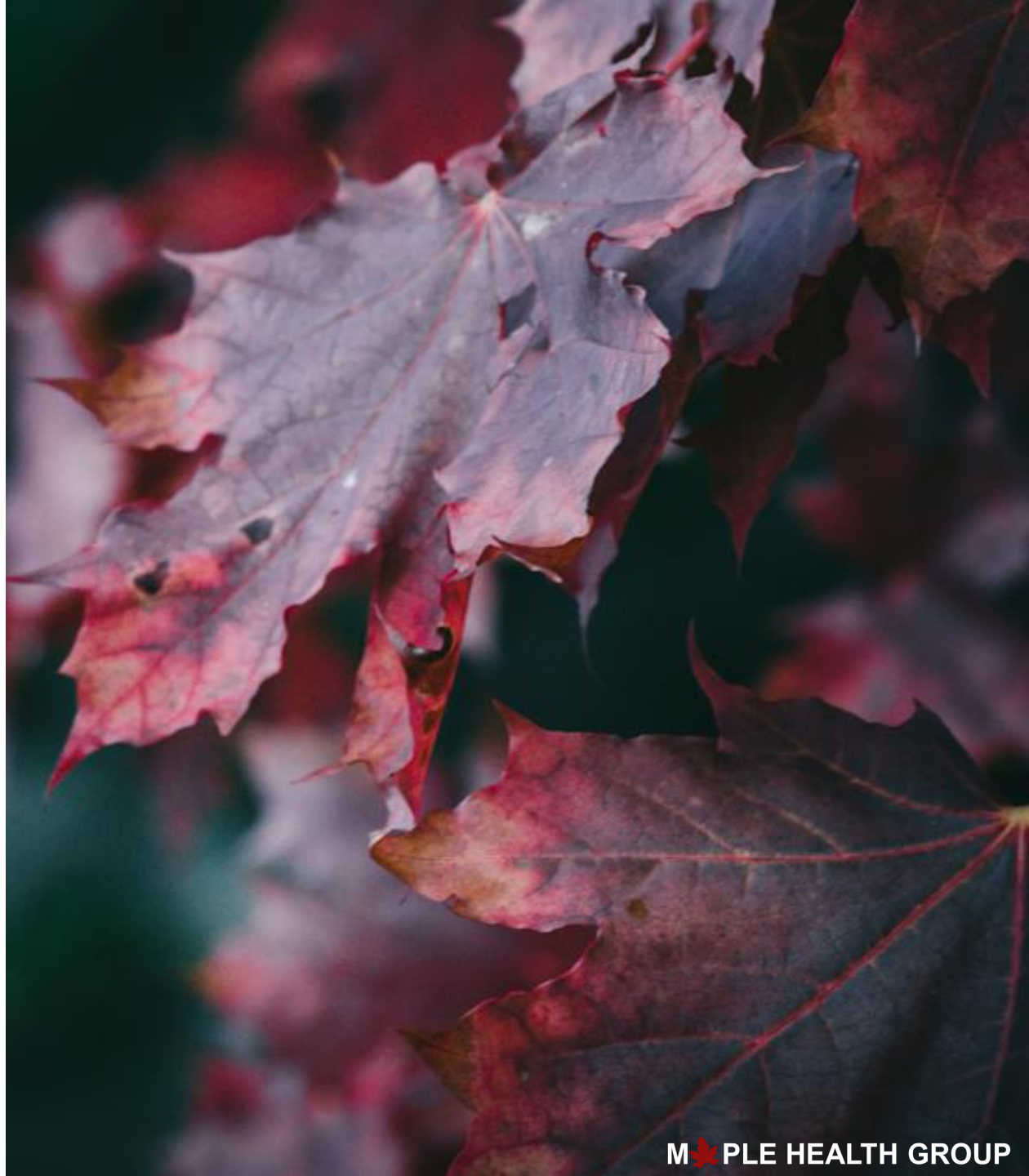
EMBED HUMAN QC
Verify decision drivers, HTA commentary, methodological limitations and source accuracy

TRACK AGENTIC WORKFLOWS
Monitor advances in agentic AI workflows and evaluate their potential to improve the accuracy and consistency

NOTE: Time efficiency was not assessed during the research. Abbreviations: AI: Artificial Intelligence; GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment; QC: Quality check.

Agenda

1. RESEARCH OBJECTIVES AND APPROACH
2. KEY RESEARCH FINDINGS
3. **RESEARCH LIMITATIONS**
4. Q&A



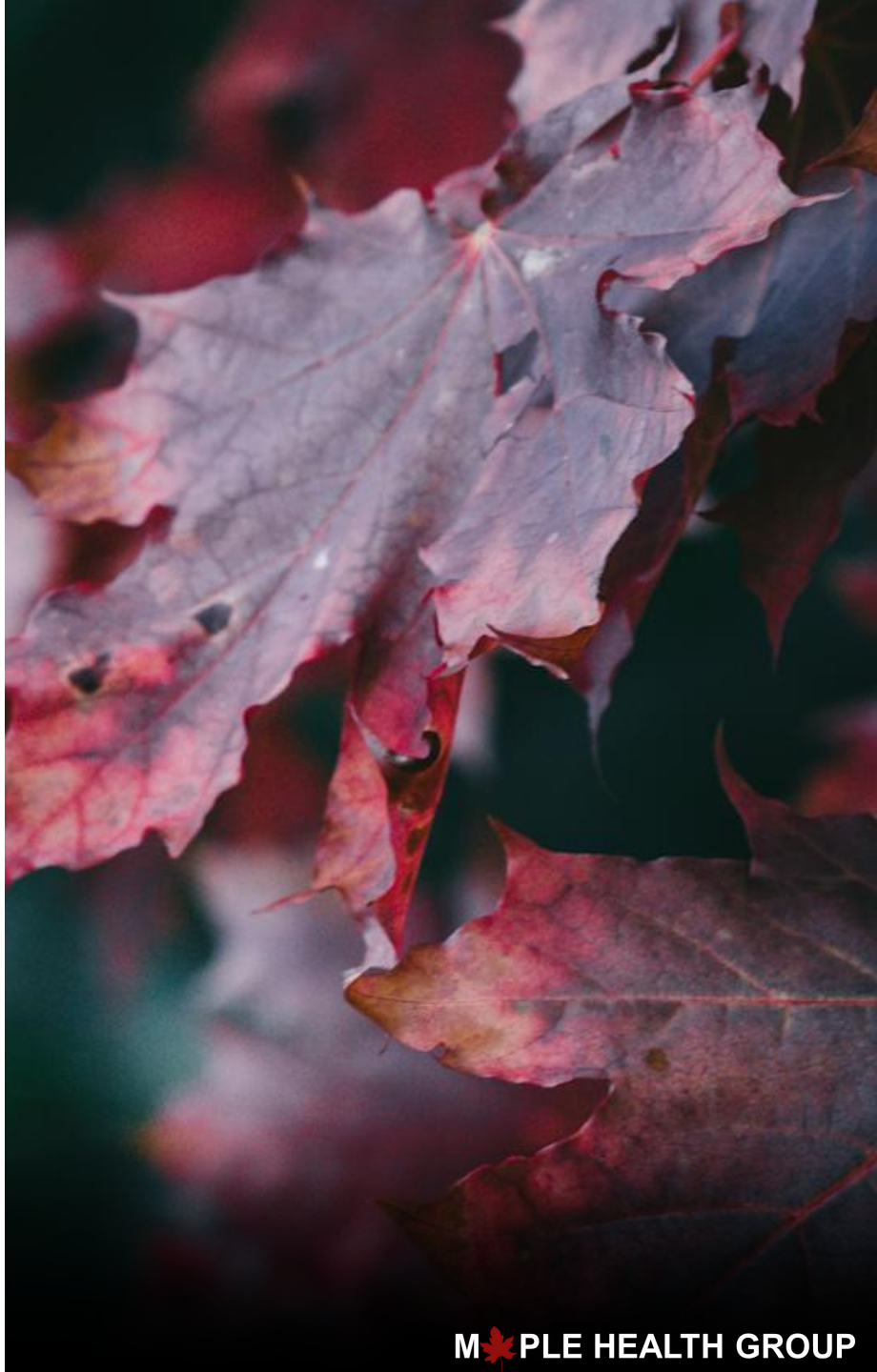
Interpreting GenAI performance: Key study limitations and considerations.

FINDING GENERALISABILITY	• Analysis of a single HTA report may not capture variations in GenAI performance across HTA agencies, reporting practices, and reports of differing complexity.
GENAI MODEL COVERAGE	• Only 3 GenAI models were evaluated, limiting applicability to other models. GenAI performance changes with model updates or new releases, affecting the applicability of findings over time.
PROMPT DEPENDENCY	• Findings may depend on prompt structure and refinement, with alternative prompting strategies potentially producing different results.
CONTROLLED RESEARCH	• Performance may be poorer without the pre-defined source document, potentially increasing the risk of hallucinations and inaccurate findings.
SCORING SUBJECTIVITY	• GenAI model accuracy was assessed by human reviewers using a subjective two-point scoring scale, introducing potential reviewer bias.
BRODAER APPLICABILITY	• Performance in a predefined extraction task may not fully reflect GenAI's ability to support broader, more complex HTA research workflows.

NOTE: The research findings were not cross-validated against findings from other studies and with agentic workflows. Additionally, regarding the reproducibility of findings, conducting only two runs per model provides a limited indication of output consistency and variability. Abbreviations: GenAI: Generative Artificial Intelligence; HTA: Health Technology Assessment.

THANK YOU FOR YOUR ATTENTION.

ANY QUESTIONS?



For more information, please contact us.

MAPLE HEALTH GROUP

Rachel Beckerman, PhD
Managing Partner & Founder

Email : rachel.beckerman@maplehealthgroup.com
Mob : (647) 242-8073

Dennis Meletiche, PharmD
Managing Partner

Email : dennis.meletiche@maplehealthgroup.com
Mob : (617) 823-5996

Nate Smith, PhD
Managing Partner

Email : nate.smith@maplehealthgroup.com
US Mob : (203) 676-3941
IT Mob : +39 333-3564245

Rob Blissett, PhD
Partner

Email : rob.blissett@maplehealthgroup.com
Mob : +44 (0) 7917-352-141

Maple Health Group, LLC
Health Economics and Market Access Consulting

website: www.maplehealthgroup.com