WHAT IS THE VALUE OF LARGE LANGUAGE MODELS FOR IMPROVING EXTRACTION AND INTERPRETATION OF EFFICIENCY DATA FROM THE FRENCH COMMISSION D'ÉVALUATION ÉCONOMIQUE ET DE SANTÉ PUBLIQUE (CEESP): DECISIONS ON BREAST CANCER MEDICATIONS?

MSR221

Lamarsalle Ludovic[1], Prodel Martin[2], Roze Stéphane[3]

[1]Healstra, Lyon, France | [2]DALI, Lyon, France | [3]HOMAX ADVISORY, Pommiers, France |

Contact Information:
Ludovic LAMARSALLE, MSc, PharmD
llamarsalle@healstra.com

MyHA

INTRODUCTION

- Health economic evidence interpretation is necessary for reimbursement decisions.
- French CEESP assesses products with ASMR I-III and expected sales >€20M annually in year 2.
- Manual data extraction is time-consuming and resource-intensive.
- Large Language Models (LLMs) offer potential for automation.

OBJECTIVES

Primary Objective:
Assess accuracy and reliability of three leading LLMs in extracting structured health economic data from CEESP opinions

Secondary Objectives:

- Compare inter-LLM consistency patterns
- Identify field categories with highest/lowest accuracy
- Evaluate LLM performance vs. expert validation

METHODS

Data Source:
16 CEESP opinions on breast cancer medications (march 2014 - april 2024)

LLM Models:
ChatGPT-4o
Claude-3.5-Sonnet
Mistral-Large-24.11

Extraction Protocol:
Unified prompt covering **65 data fields**:

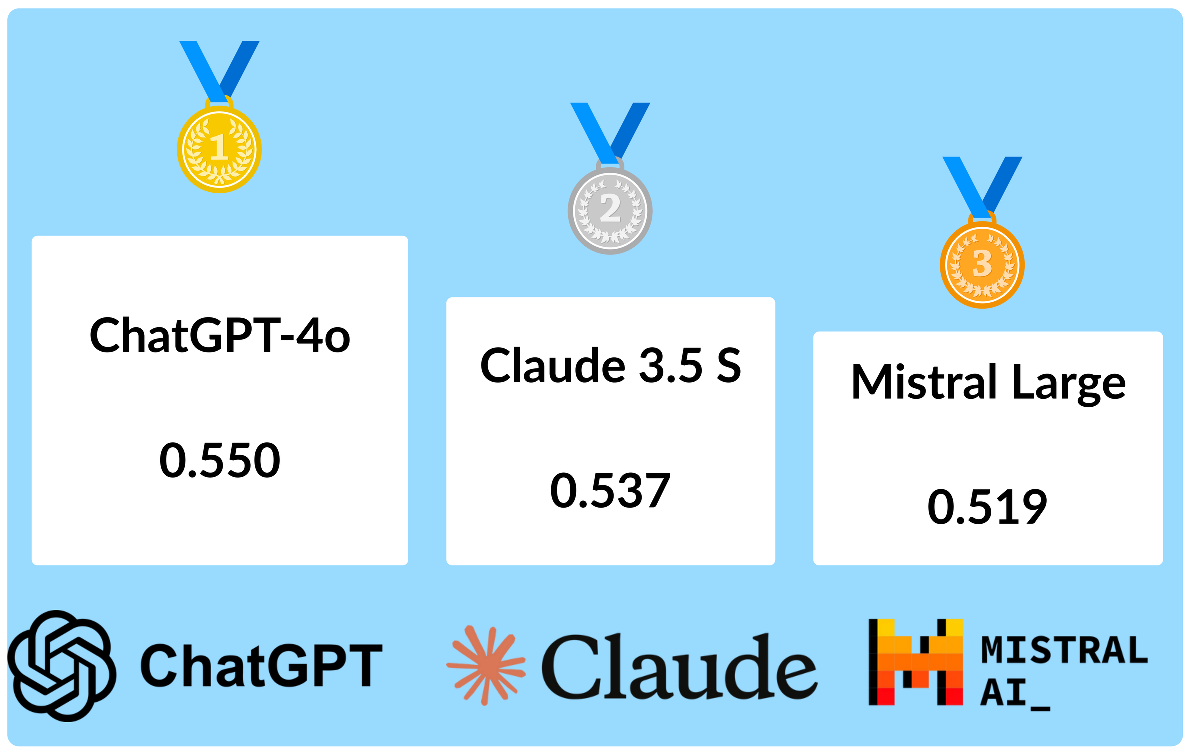
- **23** administrative variables
- **42** health-economic variables

Reference Standard:
Manual extraction by senior health economists

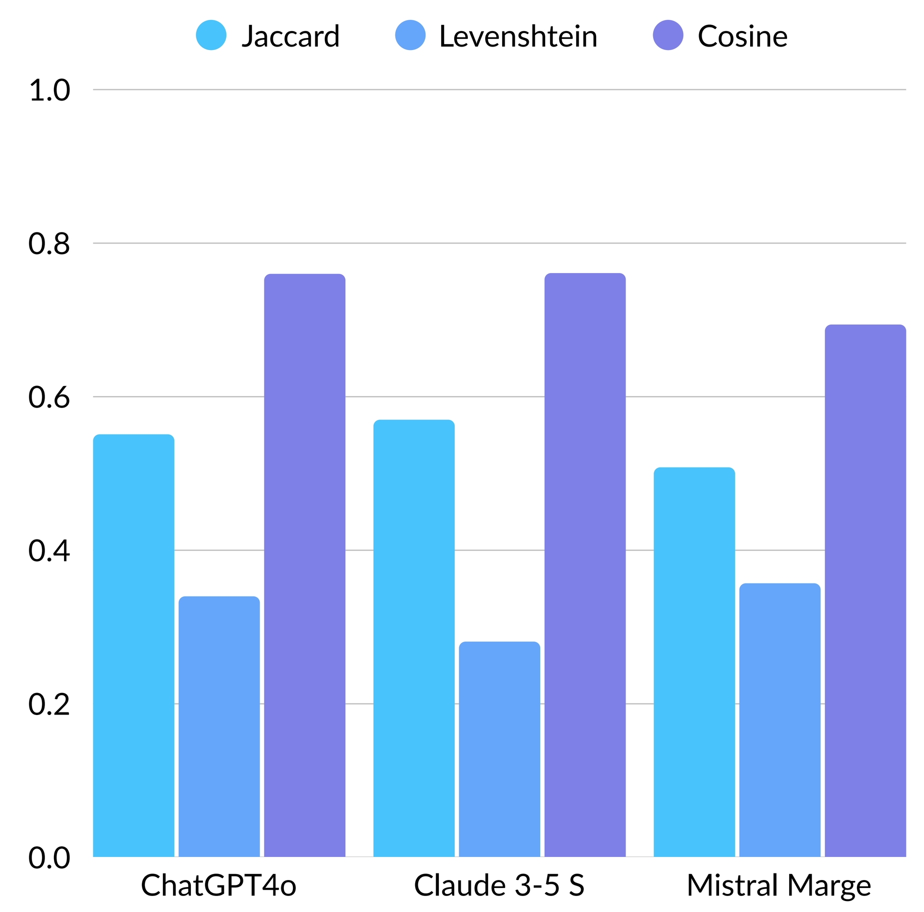
Similarity Metrics:
Jaccard Index, Levenshtein Distance, Cosine Similarity

RESULTS - LLM PERFORMANCES

Mean similarity vs expert-validated reference



Detailed Similarity Metrics of LLM vs expert reference



Top 5 Best extracted Fields

1. Collective perspective	0.948
2. Lifetime horizon	0.912
3. Claimed ASMR	0.902
4. Study types	0.826
5. SMR	0.714

Top 5 Worst extracted Fields

1. Target population	0.063
2. Efficiency commentary	0.139
3. Model type	0.270
4. Mean ICER	0.326
5. Disease category	0.374

Field Level Performance

Excellent (≥90%)

Good (70% -90%)

Low (<50 %)

Key Finding:
3 LLM consensus ≠ accuracy of expert
70% of the time 2 out of 3 LLM agree on the extracted field value.
34% of the time all 3 LLM agree, but only **31%** excellent reference match
→ *a strong human tuning is required*

Practical Recommendations

1. Deploy LLMs for administrative field extraction
2. Implement multi-LLM consensus for quality control
3. Trigger expert review for disagreements and critical fields

Expected efficiency gain:

40-60% workload reduction

CONCLUSIONS

Moderate Overall Performance
All LLMs achieved moderate similarity (0.52-0.55) with significant performance gap between structured and unstructured fields

Field-Type Dependent Accuracy
✅ Excellent: Administrative data (>90%)
❌ Poor: Clinical descriptions (<40%)
❌ Critical gap: ICER extraction (29.3%)

Consensus ≠ Correctness
Inter-model consistency overestimates real-world accuracy. Expert validation remains essential.

DISCUSSION

Promise with Constraints
State-of-the-art LLMs show promising capability for automating routine data extraction, but accuracy remains insufficient for unsupervised deployment.

Selective Reliability
LLMs excel at structured administrative data (>90%) but struggle with complex medico-economic parameters (<40%).

Expert Validation Imperative
Mandatory expert oversight required for fields impacting reimbursement decisions (ICER, populations, efficiency conditions).

ISPOR Europe 2025 | 9-12 November 2025 | Glasgow, Scotland