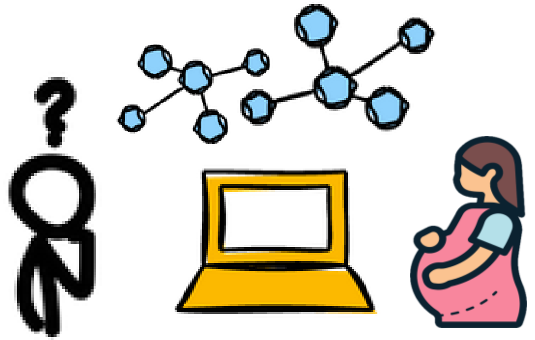
Assessing Large Language Models in Clinical Teratology: Preliminary Findings from a Multi-Model Performance Analysis

Yusuf Cem Kaplan[1*], Talha Gürsoy[2], Melih Kaan Sözmen[3], Mine Kadıoğlu Duman[2]
Barış Karadaş[1, 4]

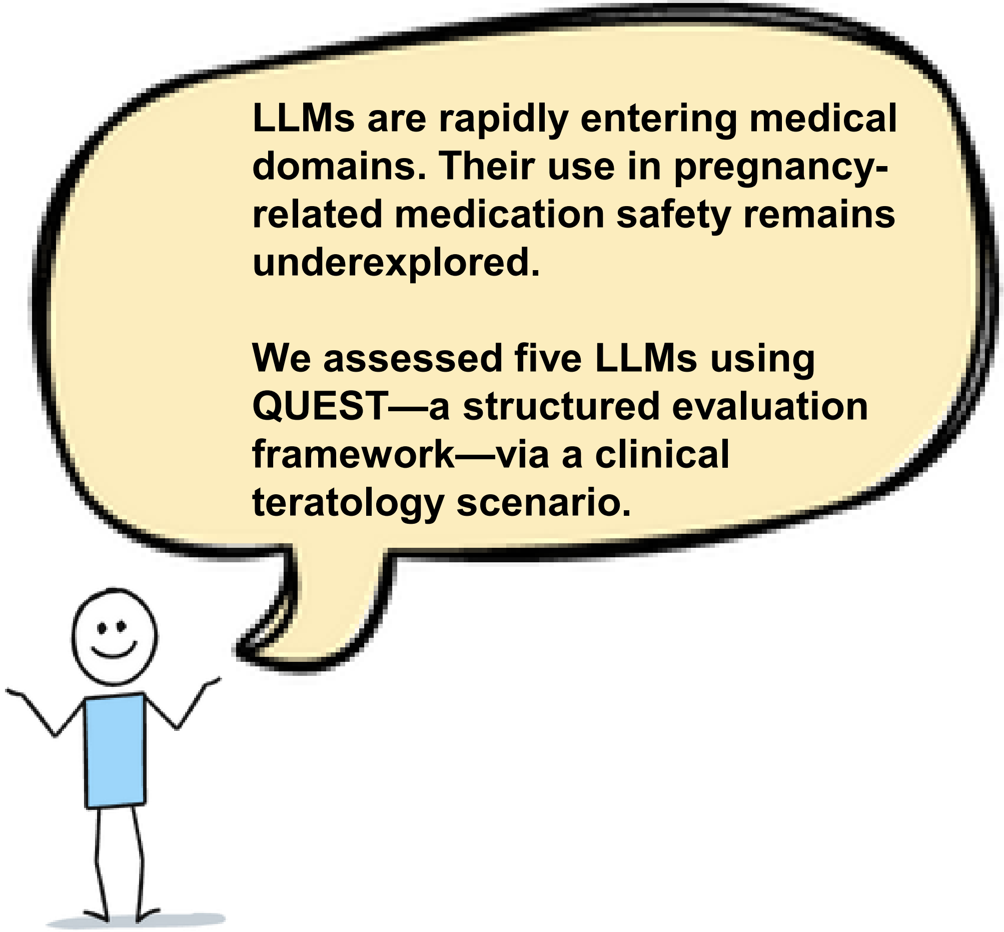
[1] *Terafar -Izmir Katip Celebi University Teratology Information, Training and Research Center, Izmir, Türkiye*
[2] *Department of Pharmacology, Karadeniz Technical University School of Medicine, Trabzon, Türkiye*
[3] *Department of Public Health, Izmir Katip Celebi University School of Medicine, Training and Research Center, Izmir, Türkiye*
[4] *Department of Pharmacology, Izmir Kâtip Celebi University School of Medicine, Izmir, Türkiye*



Terafar

İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

Introduction:



LLMs are rapidly entering medical domains. Their use in pregnancy-related medication safety remains underexplored.

We assessed five LLMs using QUEST—a structured evaluation framework—via a clinical teratology scenario.

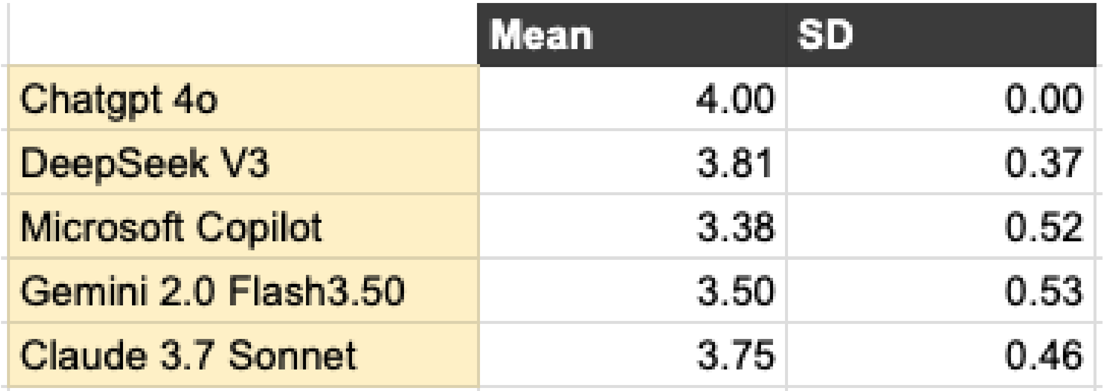
Methods:

- Scenario: 7-week pregnant user, exposed to doxylamine/pyridoxine.
- Models: ChatGPT-4o, Gemini 2.0 Flash, Claude 3.7, DeepSeek V3, Copilot.
- Evaluation: QUEST-based 8-item scale, 1–4 Likert, rated by 2 teratologists (>15 yrs exp.)
- Reliability: Cohen’s Kappa & Gwet’s AC.

ChatGPT deepseek Copilot

Gemini Claude

Results:



	Mean	SD
Chatgpt 4o	4.00	0.00
DeepSeek V3	3.81	0.37
Microsoft Copilot	3.38	0.52
Gemini 2.0 Flash3.50	3.50	0.53
Claude 3.7 Sonnet	3.75	0.46

ChatGPT-4o scored highest: 4.00 ± 0.00

DeepSeek next best: 3.81 ± 0.37

Lowest: Microsoft Copilot (3.38 ± 0.52)

Reviewer agreement:
Cohen's Kappa = 0.403 (moderate)
Gwet's AC = 0.684 (substantial)

Conclusions:

ChatGPT-4o and DeepSeek gave higher-quality responses.

Notable differences between models.

Reviewer agreement was only moderate → subjectivity matters.

Structured, domain-specific validation frameworks are needed.