

Claude Opus 4.6 interpreted urodynamic studies with **near-perfect agreement** with expert urologists ($\kappa = 0.885$)

...but its rare errors were clinically relevant.

Agreement between large language models and expert urologists in structured interpretation of urodynamic studies

Gutiérrez-Bardallo, A; Mazuecos-Quirós, J; García Sánchez, C Department of Urology, Hospital Universitario Virgen del Rocío, Seville, Spain

www.ics.org/2026/abstract/

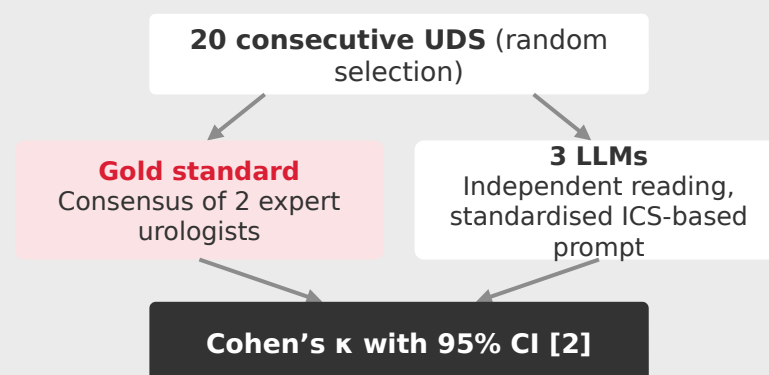
560

Introduction & aim

Urodynamic (UDS) interpretation requires expertise and ICS terminology [1]. LLMs perform well in medical tasks, but structured UDS interpretation has not been evaluated.

Aim: compare Claude Opus 4.6, DeepSeek V3 and Gemini 3 Pro with an expert panel, and judge whether discordances would change treatment.

Methods



Variables compared

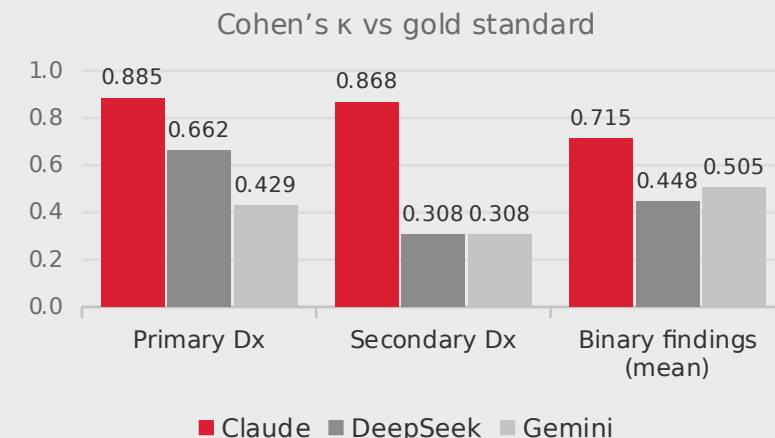
- Primary Dx (answers the urodynamic question and guides treatment) and secondary Dx, 9 categories
- 14 binary UDS findings + leak type (4 categories)
- Clinical impact of each primary-Dx discordance, judged by the experts

Interpretation

Claude's only 2 errors were in classically difficult distinctions (BOO vs acontractility; subtle low compliance), where inter-observer variability among urologists is also high.

DeepSeek and Gemini struggled to integrate simultaneous findings (45% agreement on secondary Dx). N = 20 limits precision.

Results



Overall agreement

90.6%

Claude

79.1%

DeepSeek

77.6%

Gemini

Claude: near-perfect agreement on both diagnoses; $\kappa \geq 0.81$ in 5/14 and ≥ 0.61 in 11/14 binary findings.

Only 2 primary-Dx discordances – BOO read as detrusor underactivity, and low compliance read as normal – both clinically significant (de-obstructive surgery vs catheterisation; treatment vs discharge).

Conclusions

LLMs, particularly Claude, can reach near-perfect agreement with experts in structured UDS interpretation – a potential decision-support tool, **always under specialist supervision**. Larger multicentre studies are needed.

Abbreviations: UDS, urodynamic study; LLM, large language model; Dx, diagnosis; κ , Cohen's kappa; BOO, bladder outlet obstruction.

[1] Haylen BT, et al. IUGA/ICS joint report on the terminology for female pelvic floor dysfunction. Neurourol Urodyn 2010;29:4–20. [2] Landis JR, Koch GG. Biometrics 1977;33:159–174.