Abstract Number: 10951

JOURNAL OF THE JFA Foot & Ankle

# How do ChatGPT and other generative artificial intelligence models perform on foot and ankle questions from the Brazilian Orthopedics and Traumatology Association's TEOT and TARO exams? The implications of large language models for medical education

Daniel Soares Baumfeld[1], Roberto Zambelli de Almeida Pinto[2], Paula Costa Machado[1], Lucca Gontijo Giarola[3], Lacerda MCR[4]

1. Hospital Felício Rocho, Belo Horizonte, MG, Brazil
2. Hospital Mater Dei, Belo Horizonte, MG, Brazil
3. Hospital das Clínicas da UFMG, Belo Horizonte, MG, Brazil
4. Faculdade de Ciências Médicas, Belo Horizonte, MG, Brazil

**Correspondence:** Paula Costa Machado **Email:** paulacostamachadoo@gmail.com

**Introduction:** Generative artificial intelligence (AI) is increasingly used for study and rapid consultation. We assessed how leading large language models (LLMs) perform on Brazilian Orthopedics and Traumatology Association (SBOT) Foot and Ankle exam questions.

**Methods:** Cross-sectional benchmarking of 107 foot and ankle questions from TEOT and TARO exams. Items were classified into the following categories: adult trauma, pediatric trauma, anatomy/imaging, physical examination, congenital/pediatric disorders, and adult disorders. Four generative AI models were queried with standardized prompts; responses were scored against the official key. Outcome: overall accuracy.

**Results:** ChatGPT (GPT-5 Thinking) had the highest accuracy (86.91%), followed by Gemini (79.43%). Accuracy differed by domain, with lower performance in pediatric trauma and congenital disorders. No model achieved perfect agreement with the key.

**Conclusions:** Popular generative AI models performed well on SBOT foot and ankle exam questions, with ChatGPT (GPT-5 Thinking) scoring highest. LLMs may be helpful adjuncts in residency education when used with supervision and critical appraisal.

**Keywords:** Medical education; Orthopedics; Foot.

DOI: https://doi.org/10.30795/jfootankle.2026.v20.2051

**This abstract was presented at the XXII Brazilian F&A Meeting 2026, held in São Paulo, Brazil, from April 18 to 21, 2026.**