บทวิทยาการ

การประเมินคำตอบของแชทบอทภาษาไทยสำหรับผู้ป่วยที่ได้รับการผ่าตัดขากรรไกรร่วมกับการจัดฟัน

Assessment of Thai-Language Chatbot Responses for Patients Regarding Orthognathic Surgery

อรรถพล ยงวิกุล[1,2], อัญญา วิหครัตน์[2], ณัฐรินทร์ วงศ์สิริฉัตร[3], ทองนารถ คำใจ[1,4]

Atapol Yongvikul[1,2], Anya Wihokrat[2], Nattharin Wongsirichat[3], Thongnard Kumchai[1,4]

[1]แผนกศัลยกรรมช่องปากและแม็กซิลโลเฟเชียล คณะทันตแพทยศาสตร์ มหาวิทยาลัยกรุงเทพธนบุรี กรุงเทพมหานคร ประเทศไทย

[1]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Bangkokthonburi University Bangkok, Thailand

[2]แผนกศัลยกรรมช่องปากและแม็กซิลโลเฟเชียล โรงพยาบาลศัลยกรรมมาสเตอร์พีซ กรุงเทพมหานคร ประเทศไทย

[2]Department of Oral and Maxillofacial Surgery, Masterpiece Plastic Hospital, Bangkok, Thailand

[3]แผนกทันตกรรมจัดฟัน คณะทันตแพทยศาสตร์ มหาวิทยาลัยกรุงเทพธนบุรี กรุงเทพมหานคร ประเทศไทย

[3]Department of Orthodontics, Faculty of Dentistry, Bangkokthonburi University Bangkok, Thailand

[4]สำนักคณบดี คณะทันตแพทยศาสตร์ มหาวิทยาลัยกรุงเทพธนบุรี กรุงเทพมหานคร ประเทศไทย

[4]Deanery office, Faculty of Dentistry, Bangkokthonburi University Bangkok, Thailand

บทคัดย่อ

ปัญญาประดิษฐ์ (AI) โดยเฉพาะแบบจำลองภาษาขนาดใหญ่ (LLM) ได้พัฒนาอย่างรวดเร็วในช่วงไม่กี่ปีที่ผ่านมา และถูกนำไปประยุกต์ใช้ในหลายภาคส่วน รวมถึงด้านสุขภาพและทันตกรรม หนึ่งในแอปพลิเคชันที่โดดเด่นคือการพัฒนาแชตบอตที่สามารถสนทนาเลียนแบบมนุษย์ผ่านการประมวลผลภาษาธรรมชาติ (NLP) ซึ่งมีบทบาทในการให้ความรู้แก่ผู้ป่วย โดยเฉพาะการอธิบายขั้นตอนทางการแพทย์ที่ซับซ้อนให้เข้าใจง่าย อย่างไรก็ตาม คุณภาพของข้อมูลที่แชตบอตให้ในภาษาที่ไม่ใช่ภาษาอังกฤษ เช่น ภาษาไทย ยังไม่ได้รับการศึกษามากนัก งานวิจัยนี้มีวัตถุประสงค์เพื่อประเมินคุณภาพของคำตอบภาษาไทยที่สร้างโดยแชตบอตซึ่งขับเคลื่อนด้วย LLM เกี่ยวกับการผ่าตัดขากรรไกรร่วมกับการจัดฟัน (OGS) เพื่อประเมินคุณภาพของข้อมูลที่แชตบอตภาษาไทยซึ่งขับเคลื่อนด้วย LLM ให้แก่คำถามที่พบบ่อยจากผู้ป่วยเกี่ยวกับการผ่าตัดขากรรไกรร่วมกับการจัดฟัน ศัลยแพทย์ช่องปากและแม็กซิลโลเฟเชียลที่ได้รับการรับรองจำนวน 2 คน ได้จัดทำชุดคำถามที่พบบ่อยจำนวน 10 ข้อในภาษาไทยเกี่ยวกับ OGS และนำส่งไปยังแพลตฟอร์มแชตบอต AI จำนวน 4 แห่ง ได้แก่ ChatGPT (GPT-4o), Google Gemini (Gemini 2.0 Flash), Anthropic Claude (Claude 3.5 Sonnet) และ Microsoft Copilot (เวอร์ชันมาตรฐานที่ใช้เหตุผลแบบ o1) โดยใช้เฉพาะเวอร์ชันฟรีหรือพื้นฐาน คำตอบจากแชตบอตถูกทำให้ไม่ระบุตัวตนและเข้ารหัสเพื่อปิดบังข้อมูลจากผู้ประเมิน ผู้เชี่ยวชาญด้าน OGS จำนวน 2 คนได้ประเมินคำตอบแต่ละชุดอย่างอิสระโดยใช้ Global Quality Score (GQS) ซึ่งครอบคลุม 5 ด้าน ได้แก่ ความถูกต้อง ความครบถ้วน ความชัดเจน ความเกี่ยวข้อง และความสม่ำเสมอ คะแนนถูกบันทึกใน Microsoft Excel และวิเคราะห์ด้วยสถิติเชิงพรรณนา ความเชื่อมั่นระหว่างผู้ประเมินวัดด้วย intraclass correlation coefficient พบว่า Gemini 2.0 Flash ได้คะแนนสูงสุดในด้านความถูกต้อง (3.90 ± 1.66) รองลงมาคือ Copilot (3.40 ± 1.58), GPT-4o (2.80 ± 1.23) และ Claude 3.5 Sonnet (2.40 ± 1.07) ด้านความครบถ้วน Gemini ได้คะแนนสูงสุด (4.70 ± 0.95) ขณะที่ Claude ได้คะแนนต่ำสุด (3.90 ± 1.10) ด้านความชัดเจน Copilot ได้คะแนนสูงสุด (4.50 ± 0.85) และ Claude ต่ำสุด (2.80 ± 0.79) ด้านความเกี่ยวข้อง Gemini ได้คะแนนสูงสุด (4.40 ± 0.84) และ Claude ต่ำสุด (3.30 ± 0.67) ด้านความสม่ำเสมอ Copilot ได้คะแนนสูงสุด (4.70 ± 0.48) และ Claude ต่ำสุด (3.80 ± 0.92) ความเชื่อมั่นระหว่างผู้ประเมินโดยรวมอยู่ในระดับดี คุณภาพของคำตอบที่สร้างโดยแชตบอตภาษาไทยซึ่งขับเคลื่อนด้วย LLM ในหัวข้อเกี่ยวกับ OGS มีความแตกต่างกันระหว่างแพลตฟอร์ม โดย Gemini 2.0 Flash แสดงผลการประเมินโดยรวมดีที่สุดในหลายด้าน แม้ GPT-4o จะให้เนื้อหาที่เข้าใจง่ายสำหรับผู้ใช้ทั่วไป เพื่อเพิ่มความน่าเชื่อถือของแชตบอต บุคลากรทางการแพทย์ไทย

ควรมีส่วนร่วมในการสร้างเนื้อหาภาษาไทยที่มีคุณภาพและเข้าถึงได้สำหรับการฝึกโมเดล แชตบอตสามารถเป็นเครื่องมือเสริมในการให้ความรู้แก่ผู้ป่วย แต่ไม่ควรใช้แทนการปรึกษาแพทย์ผู้เชี่ยวชาญ

คำสำคัญ: จัดฟัน, แชทบอท, ปัญญาประดิษฐ์, ผ่าตัดขากรรไกร, โมเดลภาษาขนาดใหญ่

Abstract

In recent years, artificial intelligence (AI), particularly large language models (LLM), has advanced rapidly and been applied across various sectors, including healthcare and dentistry. A prominent application is the development of chatbots that simulate human-like conversations using natural language processing (NLP). These tools can assist in patient education, especially when providing accessible explanations of complex medical procedures. However, the quality of information they provide in non-English languages, such as Thai, remains underexplored. This study aimed to assess the quality of Thai-language responses generated by LLM-based chatbots regarding orthognathic surgery combined with orthodontic treatment (OGS). To evaluate the quality of information provided by Thai-language LLM-powered chatbots in response to frequently asked patient questions about orthognathic surgery with orthodontic treatment. Two board-certified oral and maxillofacial surgeons created a set of 10 frequently asked questions in Thai about OGS. These were submitted to four major AI chatbot platforms: ChatGPT (GPT-4o), Google Gemini (Gemini 2.0 Flash), Anthropic Claude (Claude 3.5 Sonnet), and Microsoft Copilot (standard version with o1 reasoning), using only their free or basic versions. Chatbot responses were anonymized and coded to blind the evaluators. Two experienced OGS specialists independently assessed each answer using the Global Quality Score (GQS), which evaluates five domains: accuracy, completeness, clarity, relevance, and consistency. Scores were recorded in Microsoft Excel and analyzed using descriptive statistics. Inter-rater reliability was measured using the intraclass correlation coefficient. Gemini 2.0 Flash scored highest in accuracy (3.90 ± 1.66), followed by Copilot (3.40 ± 1.58), GPT-4o (2.80 ± 1.23), and Claude 3.5 Sonnet (2.40 ± 1.07). For completeness, Gemini led again (4.70 ± 0.95), while Claude had the lowest score (3.90 ± 1.10). In clarity, Copilot ranked highest (4.50 ± 0.85), and Claude lowest (2.80 ± 0.79). In relevance, Gemini scored highest (4.40 ± 0.84), while Claude trailed (3.30 ± 0.67). Copilot achieved the highest consistency (4.70 ± 0.48), and Claude the lowest (3.80 ± 0.92). The overall inter-rater reliability for GQS was good. The quality of responses generated by Thai-language LLM chatbots on OGS-related topics varied between platforms. Google Gemini 2.0 Flash demonstrated the highest overall performance across multiple evaluation domains. While GPT-4o produced understandable content for general users. To enhance chatbot reliability, Thai healthcare professionals are encouraged to contribute high-quality, accessible Thai-language content for model training. Chatbots may serve as supplementary tools for patient education but should not replace professional medical consultation.

Keyword: Orthognathic Surgery, Chatbot, Artificial Intelligence, Orthodontics treatment, Large Language Model

Received date: Aug 15, 2025 **Revised date:** Oct 24, 2025 **Accepted date:** Nov 5, 2025
Doi: 10.14456/jdat.2026.5

ติดต่อเกี่ยวกับบทความ:
ณัฐรินทร์ วงศ์สิริฉัตร แผนกทันตกรรมจัดฟัน คณะทันตแพทยศาสตร์ มหาวิทยาลัยกรุงเทพธนบุรี 16/10 ถ.เรียบคลองทวีวัฒนา เขต/แขวงทวีวัฒนา กรุงเทพมหานคร 10170 ประเทศไทย โทร: 02-4315383 อีเมล: nattharin_wong@gmail.com
Correspondence to:
Nattharin Wongsirichat, Department of Orthodontics, Faculty of Dentistry, Bangkokthonburi University, 16/10 Thawi Watthana, Bangkok, 10170, Thailand. Tel: 02-4315383 Email: nattharin_wong@gmail.com

บทนำ

การผ่าตัดเลื่อนขากรรไกร (orthognathic surgery: OGS) ร่วมกับการจัดฟัน (Orthodontic treatment) เป็นกระบวนการรักษาที่ซับซ้อนและใช้ระยะเวลานาน รวมถึงต้องอาศัยความร่วมมืออย่างใกล้ชิดระหว่างทันตแพทย์จัดฟัน ศัลยแพทย์ช่องปากและแม็กซิลโลเฟเชียล รวมถึงแพทย์ และทันตแพทย์ทั่วไป[1,2] ผู้ป่วยมักดำเนินการเสาะแสวงหาความรู้และคำแนะนำทางการแพทย์จากผู้เชี่ยวชาญเมื่อต้องตัดสินใจเกี่ยวกับแนวทางการรักษา และศึกษาข้อมูลเกี่ยวกับขั้นตอนก่อนและหลังการผ่าตัด ตลอดจนความเสี่ยงที่อาจเกิดขึ้น อย่างไรก็ตาม ความไม่สะดวกในการเข้าถึงผู้เชี่ยวชาญทางสุขภาพตลอดเวลาที่ต้องการ ประกอบกับความสนใจของผู้ป่วยที่จะเรียนรู้จากประสบการณ์ของผู้อื่นที่เคยผ่านการผ่าตัดคล้ายกัน ส่งผลให้ผู้ป่วยจำนวนไม่น้อยหันไปค้นคว้าข้อมูลจากแหล่งต่างๆ ในระบบสารสนเทศน์เช่น อินเทอร์เน็ต[3,4] ซึ่งในปัจจุบันมีสารสนเทศน์ออนไลน์หลายรูปแบบที่ให้ข้อมูลเกี่ยวกับการ OGS ทั้งในรูปแบบข้อความและสื่อภาพ[5-8] แม้ว่าวรรณกรรมทางวิชาการที่ยึดหลักฐานทางวิทยาศาสตร์จะเป็นแหล่งข้อมูลที่เชื่อถือได้มากที่สุด แต่เนื้อหาส่วนใหญ่มักมุ่งเน้นสื่อสารกันในระหว่างผู้เชี่ยวชาญ และเข้าถึงไม่บ่อยนักโดยตัวผู้ป่วยเอง ทั้งที่ควรเป็นบุคคลผู้ได้รับข้อมูลที่ถูกต้องครบถ้วนมากที่สุด อนึ่ง ทั้งยังมีแหล่งข้อมูลอีกประเภทหนึ่งที่เรียกว่าวรรณกรรมสีเทา (grey literature) ซึ่งเป็นข้อมูลที่ไม่มีรูปแบบการนำเสนอที่เป็นมาตรฐานและขาดกระบวนการทบทวนโดยผู้ทรงคุณวุฒิ ส่งผลให้เนื้อหาในแหล่งข้อมูลดังกล่าวอาจคลาดเคลื่อนหรือไม่ครบถ้วน[9,10] อีกช่องทางสำคัญที่ผู้เตรียมเข้ารับการทำ OGS ใช้ในการค้นหาข้อมูลคือสื่อออนไลน์และแพลตฟอร์มต่างๆ เช่น เว็บไซต์คลินิก, แพลตฟอร์มวิดีโอ (YouTube), และสื่อสังคมออนไลน์ ได้แก่ Instagram, Facebook, Bing, เว็บบอร์ดสนทนา ซึ่งผู้ป่วยมักเข้าเยี่ยมชมเพื่อค้นคว้าเกี่ยวกับกระบวนการผ่าตัด อย่างไรก็ตาม การศึกษาประเมินคุณภาพของข้อมูลที่เผยแพร่บนช่องทางเหล่านี้พบว่ามีความ แตกต่างและผันผวนด้านคุณภาพอย่างมีนัยสำคัญ โดยเฉพาะเว็บไซต์ทั่วไป[4] และเนื้อหาบนแพลตฟอร์มวิดีโอ[5] รวมถึงเว็บบอร์ดสนทนา[4,6,7,11] มักมีเนื้อหาที่ต่ำกว่ามาตรฐาน[5] หรือมีการสอดแทรกอารมณ์และความเห็นส่วนบุคคลเช่น การรีวิวจากผู้ป่วยโดยตรง[4,6] ซึ่งอาจทำให้ความถูกต้องแม่นยำของข้อมูลลดลง[6] ดังนั้นจึงมีความจำเป็นที่ บุคลากรทางการแพทย์ควรเข้ามามีบทบาทในการชี้แนะผู้ป่วยให้เข้าถึงแหล่งข้อมูลออนไลน์ที่เหมาะสม[6] รวมถึงชี้ให้เห็นว่าแม้ข้อมูลจากอินเทอร์เน็ตจะเข้าถึงได้สะดวก แต่ผู้ป่วยควรตระหนักถึงความหลากหลายของคุณภาพเนื้อหาและรูปแบบการนำเสนอเพื่อหลีกเลี่ยงการรับข้อมูลที่ผิดพลาดอันเนื่องมาจากการเผยแพร่จากผู้ใช้งานทั่วไปที่ขาดความรู้ทางการแพทย์ที่ถูกต้อง[6]

ในช่วงไม่กี่ปีที่ผ่านมาเทคโนโลยีปัญญาประดิษฐ์ (artificial intelligence: AI) ได้เติบโตอย่างก้าวกระโดด โดยเฉพาะเทคโนโลยีที่เกี่ยวข้องกับการเรียนรู้ของmachine learning และแบบจำลองภาษาขนาดใหญ่ (large language models: LLM) ซึ่งถูกนำมาใช้ในหลากหลายสาขา รวมถึงทันตกรรมด้วย หนึ่งในแอปพลิเคชันที่สำคัญของ AI คือแชทบอท (AI Chatbot:AICB) ซึ่งสามารถทำความเข้าใจ วิเคราะห์ และตอบสนองต่อคำถามของมนุษย์ได้อย่างเป็นธรรมชาติผ่านกระบวนการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ซึ่งเป็นแขนงวิชาหนึ่งของ AI[12,13] โดยในปัจจุบัน วิธีการทาง NLP ที่ทันสมัยอาศัยแบบจำลอง LLM เป็นหลัก โดยแบบจำลองเหล่านี้ได้รับการฝึกฝนจากชุดข้อมูลขนาดใหญ่ และสามารถอธิบายการแจกแจงเชิงสถิติของคำ ภาพ อักขระและเครื่องหมายวรรคตอนที่ปรากฏในข้อความที่มนุษย์ได้สร้างขึ้นและเผยแพร่สู่สาธารณะ[13] LLM เป็นระบบที่สามารถตั้งโปรแกรมเพื่อสร้างข้อความที่ลื่นไหลและสอดคล้องกัน ตอบคำถาม แปลภาษา และดำเนินงานที่เกี่ยวข้องกับภาษาได้หลากหลายรูปแบบเสมือนหนึ่งว่าถูกตอบคำถามโดยมนุษย์[14] การประยุกต์ใช้เทคโนโลยีที่ขับเคลื่อนด้วยAIเหล่านี้มาเป็นส่วนช่วยในงานสาขาทันตกรรม ส่งผลให้เกิดการเพิ่มขึ้นของศักยภาพในด้านต่างๆ ไม่ว่าจะเป็นการวินิจฉัยโรค การวางแผนการรักษา และกระบวนการทางคลินิก โดยเฉพาะโปรแกรม AICB ที่ขับเคลื่อนด้วย LLM ซึ่งพัฒนาขึ้นบนพื้นฐานของการปฏิสัมพันธ์ระหว่างมนุษย์กับคอมพิวเตอร์อย่างชาญฉลาด โดยมีเป้าหมายเพื่อเลียนแบบบทสนทนาธรรมชาติที่ตอบสนองต่อผู้ใช้งานในสภาพแวดล้อมออนไลน์ได้อย่างเหมาะสมและมีประสิทธิภาพ[13,15]

จากข้อมูลข้างต้นจะเห็นได้ว่า ผู้ป่วยกำลังจะเข้ารับ อยู่ในกระบวนการ หรือกระทั่งหลังรับการผ่าตัด OGS มักต้องการสืบหาข้อมูลที่ถูกต้อง ชัดเจน และเข้าถึงได้ง่ายเพื่อนำมาใช้ประกอบการตัดสินใจด้านการรักษา แม้จะมีแหล่งข้อมูลออนไลน์มากมาย ทั้งจากเว็บไซต์คลินิก วิดีโอ โซเชียลมีเดีย และเว็บบอร์ดผู้ป่วย แต่การค้นหาข้อมูลหนึ่งที่ถูกใช้อย่างแพร่หลายในปัจจุบันคือการใช้ AICB แม้ว่าเทคโนโลยี AICB ที่ใช้กระบวนการ NLP ถูกนำมาใช้เพิ่มศักยภาพในการสื่อสารข้อมูลทางทันตกรรม แต่ปัจจุบันยังไม่มีการศึกษาใดที่ประเมินประสิทธิภาพของ AICB โดยเฉพาะอย่างยิ่งด้วยภาษาท้องถิ่นคือภาษาไทยซึ่งเป็นภาษาหลักและภาษาราชการของประเทศไทยในการให้ข้อมูลเกี่ยวกับ OGS แก่ผู้ป่วยโดยตรง ดังนั้นการศึกษานี้จึงมีวัตถุประสงค์เพื่อประเมินคุณภาพ ความถูกต้อง และความเหมาะสมของข้อมูลที่ AICB ภาษาไทยให้แก่ผู้ป่วยในบริบทของการให้ข้อมูลเกี่ยวกับ OGS เพื่อสนับสนุนการตัดสินใจและเพิ่มความเข้าใจของผู้ป่วยอย่างมีประสิทธิภาพ

วัสดุอุปกรณ์และวิธีการศึกษา

เพื่อประเมินศักยภาพในการใช้แบบจำลองภาษาที่ขับเคลื่อนด้วยปัญญาประดิษฐ์ (AI-based language models) หรือ AICB ในการให้ข้อมูลแก่ผู้ป่วย แผนผังแสดงขั้นตอนของการศึกษาวิจัยนี้แสดงไว้ในรูปที่ 1 โดยการศึกษาครั้งนี้ไม่จำเป็นต้องขอการรับรองด้านจริยธรรมเนื่องจากไม่มีการใช้ข้อมูลหรือวัสดุที่ได้จากมนุษย์หรือสัตว์ในการวิจัยแต่อย่างใด กล่าวโดยสรุปการสร้างคำถามที่มักจะถูกถาม (frequent asked questions) ในการทำ OGS ถูกสร้างโดยผู้วิจัย 2 ราย (A.Y.,N.W.) ที่มีประสบการณ์ด้าน OGS รวมทั้งหมด 10 คำถามหลัก 17 คำถามย่อยในภาษาไทย (ตารางที่ 1) เพื่อให้มั่นใจว่าคำตอบที่ไม่เหมาะสม ไม่สมบูรณ์ หรือไม่ถูกต้องไม่ได้เกิดจากการออกแบบ Prompt หรือคำถามที่ไม่ดี ผู้เขียนได้ดำเนินกระบวนการตรวจสอบหลายขั้นตอน เริ่มจากการที่นักวิจัยทุกคนร่วมกันตรวจสอบคำถามทั้งหมดอย่างรอบคอบ เพื่อยืนยันถึงความเหมาะสม ความชัดเจน และความสอดคล้องกับวัตถุประสงค์ของการศึกษา นอกจากนี้ ยังได้ทำการตรวจสอบคำตอบเพื่อค้นหารูปแบบที่อาจบ่งชี้ถึงปัญหาที่เกิดจากคำถามมากกว่าความเข้าใจผิดของ AICB เช่นคำถามที่ ตอบไม่ตรงคำถาม หรือมีภาวะหลอน (Hallucination) พร้อมๆ กันในทุก AICB จากนั้นคำถามถูกป้อนเข้าไปใน AICB โดยหาก AICB มีให้เลือกระหว่างการสมัครสมาชิกชนิดเสียค่าบริการ (subscription) จะเลือกชนิดไม่เสียค่าบริการ (free or basic version) จำนวน 4 โปรแกรมที่ถูกใช้งานรวมกันทั่วโลกในไตรมาสที่ 2 ปี 2568 ถึง 89.58%[16] ได้แก่ ChatGPT (GPT-4o, OpenAI, San Francisco, CA) ซึ่งมีจุดเด่นด้านการประมวลผลภาษาที่ลื่นไหลแม่นยำ และสามารถตอบคำถามทางการแพทย์ได้อย่างครอบคลุมในรูปแบบภาษาที่เข้าใจง่าย, Google Gemini (Gemini 2.0 Flash, Google DeepMind, London, UK) ซึ่งเน้นการอ้างอิงข้อมูลจากการค้นหาแบบเรียลไทม์ผ่านระบบของ Google ทำให้เหมาะสำหรับคำถามที่ต้องการข้อมูลปัจจุบัน, Anthropic Claude (Claude 3.5 Sonnet, Anthropic, San Francisco, CA) ที่มีจุดเด่นในด้านการใช้ภาษาที่ถูกตรวจสอบความถูกต้องมาแล้วเท่านั้น และให้ข้อมูลเชิงลึกอย่างรอบคอบเหมาะสมกับผู้ป่วย, และ Microsoft Copilot (Standard Copilot with o1 reasoning, OpenAI & Microsoft, San Francisco, CA & Redmond, WA) ซึ่งผสานระบบดั้งเดิมของ GPT เข้ากับ Bing Search ทำให้สามารถสรุปข้อมูลพร้อมลิงก์อ้างอิง และรองรับการใช้งานร่วมกับผลิตภัณฑ์ของ Microsoft ได้อย่างลงตัว[17,18] คำตอบของแต่ละคำถามจะถูกกำกับไว้ด้วยรหัสเพื่อไม่ให้ผู้ประเมินคุณภาพของคำตอบทราบว่าเป็นคำตอบจากโปรแกรมใด(single blinded) จากนั้นผู้ประเมิน 2 ราย (A.Y.,A.W.) ซึ่งมีประสบการณ์ด้าน OGS มากกว่า 10 ปีจะทำการประเมินคุณภาพของคำตอบด้วย global quality score (GQS) แบบ Likert 5 ระดับ (five-point Likert-type rating scale)[19,20] ซึ่งถูกใช้อย่างแพร่หลายสำหรับประเมินคุณภาพโดยรวม ความเป็นประโยชน์ หรือความน่าเชื่อถือของข้อมูล โดยเฉพาะในการวิจัยด้านข้อมูลสุขภาพ เช่น การประเมินคุณภาพของวิดีโอใน YouTube หรือเนื้อหาออนไลน์[4-6] โดยระดับ 5 (ดีเยี่ยม) หมายถึงเนื้อหาทุกด้านมีคุณภาพสูงสุด ถูกต้อง ครบถ้วน ชัดเจน สอดคล้อง และเกี่ยวข้องโดยไม่มีข้อผิดพลาดเลย, ระดับ 4 (ดี) หมายถึงคุณภาพดีโดยรวม อาจมีจุดเล็กน้อยที่ไม่สมบูรณ์หรือคลุมเครือ แต่ไม่กระทบสาระสำคัญ, ระดับ 3 (ปานกลาง) หมายถึงคำตอบยังพอใช้ได้ มีข้อเด่นบางด้าน แต่มีจุดที่ขาดหรือผิดบ้าง, ระดับ 2 (ต่ำ) หมายถึงมีข้อผิดพลาดที่เห็นได้ชัด หรือขาดคุณภาพในหลายมิติ และระดับ 1 (แย่มาก) หมายถึงคำตอบล้มเหลวในการตอบโจทย์ ไม่มีความถูกต้อง ชัดเจน หรือเกี่ยวข้องเลย เป็นเนื้อหาที่ไม่สามารถใช้งานได้ในเชิงสาระ[20] เกณฑ์ในการประเมินแสดงใน Supplement data 1

หัวข้อการประเมินประกอบด้วย เกณฑ์ในการประเมิน 5 ด้านหลัก ได้แก่ ความถูกต้อง (Accuracy) คือความถูกต้องตามข้อเท็จจริงหรือหลักวิชาการของเนื้อหา ต้องไม่มีข้อมูลผิดพลาดและสอดคล้องกับแหล่งอ้างอิงที่น่าเชื่อถือ ความครบถ้วน (Completeness) คือการนำเสนอข้อมูลที่ครอบคลุมทุกประเด็นสำคัญโดยไม่ละเว้นส่วนที่จำเป็น ความชัดเจน (Clarity) คือความเข้าใจง่ายของภาษา การจัดเรียงเนื้อหาอย่างมีโครงสร้าง ทำให้ผู้อ่านไม่สับสน ความเกี่ยวข้อง (Relevance) คือการที่เนื้อหาอยู่ในขอบเขตของหัวข้อหรือวัตถุประสงค์ที่กำหนด ไม่มีเนื้อหานอกเรื่อง และความสม่ำเสมอ (Consistency) คือความคงที่ของแนวคิด การใช้คำศัพท์ และข้อมูลภายในเนื้อหาเดียวกันโดยไม่มีความขัดแย้งหรือไม่สอดคล้องกัน โดยตารางที่ 2 แสดงเกณฑ์การให้คะแนนในแต่ละด้าน คะแนนระดับคุณภาพของข้อมูลที่ถูกประเมินนั้นจะถูกนำมาบันทึกไว้ในโปรแกรม Microsoft Excel365 (Microsoft office365, Microsoft, Redmond, WA) โดยระหว่างการพิจารณาให้คะแนน ผู้ประเมินจะไม่ทราบชื่อของโปรแกรม AICB ที่เป็นผู้ผลิตคำตอบให้ (single blinded) มีการการวัดค่าความสอดคล้องระหว่างผู้ประเมิน (inter-rater reliability) จากนั้นระดับคะแนนที่ได้จาก 2 ผู้วิจัยจะนำมาคิดค่าเฉลี่ย ก่อนนำไปประมวลผลต่อไป

สถิติที่ใช้ในงานวิจัย

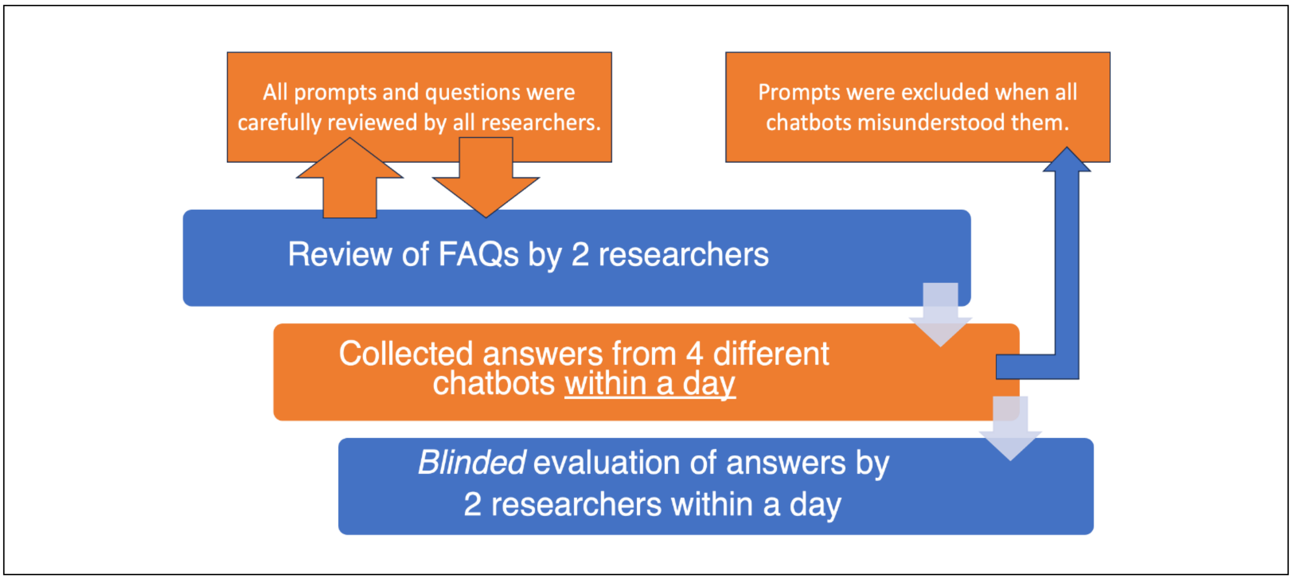
มีการคำนวณสถิติเชิงพรรณนา (descriptive statistics) ของข้อมูล ซึ่งรวมถึง ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน การเปรียบเทียบคะแนนระหว่าง 4 AICB ถูกวิเคราะห์ด้วย one-way ANOVA การวิเคราะห์ทั้งหมดในโปรแกรมวิเคราะห์สถิติ Jamovi [The jamovi project. (2025). jamovi (Version 2.6) [Computer Software]. Sydney, Australia: The jamovi project. Retrieved from https://www.jamovi.org]

ผลการศึกษา

การประเมินคุณภาพคำตอบจาก AICB ทั้งสี่โปรแกรมแสดงไว้ในตารางที่ 2 พบว่าความสอดคล้องระหว่างผู้ประเมิน (inter-rater reliability) มีค่า 0.82 ซึ่งอยู่ในระดับดี เมื่อพิจารณาผลโดยรวมในแต่ละมิติ พบว่าความสม่ำเสมอมีค่าเฉลี่ยสูงที่สุด (4.30 ± 0.79) ขณะที่ด้านความถูกต้องมีค่าเฉลี่ยต่ำที่สุด (3.13 ± 1.47) แสดงให้เห็นว่าโดยทั่วไปคำตอบจากทุกโปรแกรมมีความคงเส้นคงวาในเชิงตรรกะ แต่ยังมีความแตกต่างกันในแง่ความแม่นยำของข้อมูล เนื้อหาส่วนใหญ่มีความครบถ้วนและชัดเจนในระดับดีถึงดีมาก โดยค่าเฉลี่ยรวมของความครบถ้วนอยู่ที่ 4.18 ± 1.08 และความชัดเจนของที่ 3.73 ± 1.11 (รูปที่ 2 และตารางที่ 2) คำตอบทั้งหมดของแต่ละ AICB ถูกแสดงใน Supplement data 2

เมื่อเปรียบเทียบคุณภาพคำตอบระหว่างโปรแกรม พบว่า Gemini 2.0 Flash มีคะแนนสูงที่สุดในหลายด้าน โดยเฉพาะด้านความครบถ้วน (4.70 ± 0.95) และความถูกต้อง (3.90 ± 1.66) รองลงมาคือ Copilot with o1 reasoning ซึ่งโดดเด่นในด้านความสม่ำเสมอ (4.70 ± 0.48) ส่วน GPT-4o มีคะแนนอยู่ในระดับปานกลางในทุกมิติ ขณะที่ Claude 3.5 Sonnet ได้คะแนนต่ำที่สุดในด้านความถูกต้อง (2.40 ± 1.07) ผลลัพธ์นี้สะท้อนให้เห็นถึงความแตกต่างของศักยภาพแต่ละโมเดลในการให้ข้อมูลที่ครบถ้วน ชัดเจน และถูกต้อง โดย Gemini 2.0 Flash มีแนวโน้มให้คำตอบที่ครอบคลุมและเข้าใจง่ายที่สุดในภาพรวม (รูปที่ 2 และตารางที่ 2)

เมื่อวิเคราะห์ผลแยกตามคำถามทั้ง 10 ข้อ (ตารางที่ 1 และรูปที่ 3) พบว่า คำถามข้อ 6 และข้อ 10 ได้คะแนนความถูกต้องสูงที่สุด (4.25 คะแนน) ขณะที่ ข้อ 8 ได้คะแนนต่ำสุดเพียง 1.00 คะแนน แม้จะมีคะแนนด้านความครบถ้วนและความชัดเจนสูง แต่อาจสะท้อนถึงปัญหาความถูกต้องของข้อมูลในบางโปรแกรม ในด้านความครบถ้วน คำถามข้อ 7, 8, 9 และ 10 ได้คะแนนเต็ม 5.00 แสดงถึงการให้ข้อมูลที่ครอบคลุม ส่วนในด้านความชัดเจน คำถามข้อ 10 ได้คะแนนสูงสุด 4.50 ขณะที่ข้อ 4 และ 7 ได้คะแนนต่ำสุด 3.00 ซึ่งอาจเกี่ยวข้องกับโครงสร้างภาษาหรือรูปแบบการตอบที่ไม่ชัดเจน โดยรวมแล้ว ผลการเปรียบเทียบแสดงให้เห็นว่า Gemini 2.0 Flash มีคุณภาพของคำตอบโดยรวมดีที่สุดในทุกมิติ ขณะที่ Claude 3.5 Sonnet มีความแม่นยำต่ำที่สุด และ Copilot with o1 reasoning ให้คำตอบที่สม่ำเสมอที่สุด



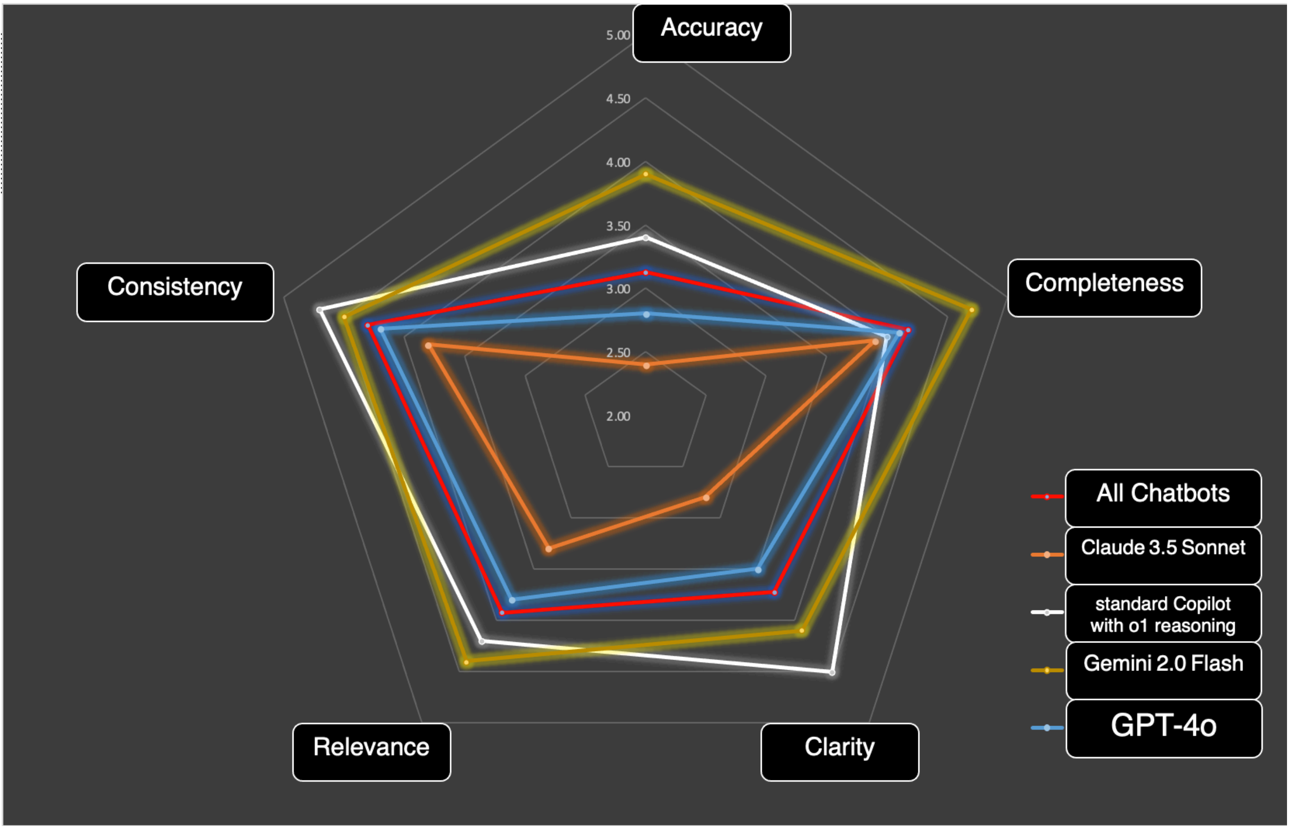
รูปที่ 1 *ภาพอธิบายกระบวนการวิจัยเกี่ยวกับการประเมินคุณภาพของคำตอบจากโปรแกรม AI chatbot 4 โปรแกรม ได้แก่ ChatGPT, Gemini, Claude และ Copilot ต่อคำถามที่พบบ่อยเกี่ยวกับ OGS ซึ่งถูกสร้างโดยผู้วิจัย 2 คน รวมทั้งหมด 10 คำถามหลัก จากนั้นใส่คำถามเข้าโปรแกรม AI โดยเลือกใช้เวอร์ชันฟรี จากนั้นบันทึกคำตอบโดยไม่เปิดเผยชื่อโปรแกรมให้ผู้ประเมินรู้ (single-blinded) ผู้ประเมินคือผู้เชี่ยวชาญ 2 รายประเมินคุณภาพคำตอบโดยใช้ Global Quality Score (GQS) แบ่งเป็น 5 ระดับซึ่งแสดงรายละเอียดในตารางที่ 2*

Figure 1 *The diagram illustrates the research process regarding the evaluation of answer quality from four AI chatbot programs: ChatGPT, Gemini, Claude, and Copilot, in response to frequently asked questions about OGS. The main questions—10 in total—were created by two researchers. These questions were then input into the AI programs using their free versions. The responses were recorded without revealing which program produced which answer (single-blinded). Two experts served as evaluators, assessing the quality of the answers using the Global Quality Score (GQS), which consists of five levels detailed in Table 2*

ตารางที่ 1 *คำถามที่ถูกถามบ่อย(frequent asked questions: FAQ) ที่ถูกสร้างโดยนักวิจัย 2 ท่าน*

Table 1 The frequently asked questions (FAQs) were created by two researchers

ชุดคำถามหลัก	จำนวนคำถามย่อย
1. การผ่าตัดขากรรไกรร่วมกับการจัดฟันคืออะไร	1
2. การผ่าตัดขากรรไกร 1 หรือ 2 ขากรรไกร แตกต่างกันอย่างไร	1
3. ข้อแตกต่างระหว่างการผ่าตัดก่อนจัดฟัน และการจัดฟันก่อนผ่าตัด สำหรับการผ่าตัดขากรรไกรร่วมกับการจัดฟัน	1
4. หลังผ่าตัดขากรรไกรร่วมกับการจัดฟัน จำเป็นต้องมัดฟันหรือไม่	1
5. การผ่าตัดขากรรไกรร่วมกับการจัดฟันควรไปพบทันตแพทย์จัดฟันก่อนหรือไปพบศัลยแพทย์แม็กซิลโลเฟเชียลก่อน เพราะอะไร	2
6. ใบหน้าเบี้ยว หรือ รอยยิ้มเบี้ยว สามารถแก้ไขโดยการผ่าตัดขากรรไกรได้หรือไม่	2
7. อาหารที่สามารถทานได้หลังจากผ่าตัดขากรรไกร และควรทานนานเท่าใด และทานอาหารปกติได้เมื่อใด	3
8. สอนการดูแลความสะอาดแผลในช่องปากหลังผ่าตัดขากรรไกร	1
9. การเลื่อนคาง จำเป็นต้องทำพร้อมกับการเลื่อนขากรรไกรหรือไม่ เพราะเหตุใด	2
10. หลังการผ่าตัดขากรรไกร มีโอกาสที่จะได้ผลลัพธ์ไม่เป็นไปตามที่ต้องการหรือไม่ หากไม่ประสบความสำเร็จ สามารถแก้ไขได้หรือไม่ ควรแก้ไขเมื่อใด	3



รูปที่ 2 *พารามีเตอร์เปรียบเทียบเกณฑ์ในการประเมิน 5 ด้านหลัก ได้แก่ ความถูกต้อง (Accuracy), ความครบถ้วน (Completeness), ความชัดเจน (Clarity), ความเกี่ยวข้อง (Relevance), และความสม่ำเสมอ (Consistency)*

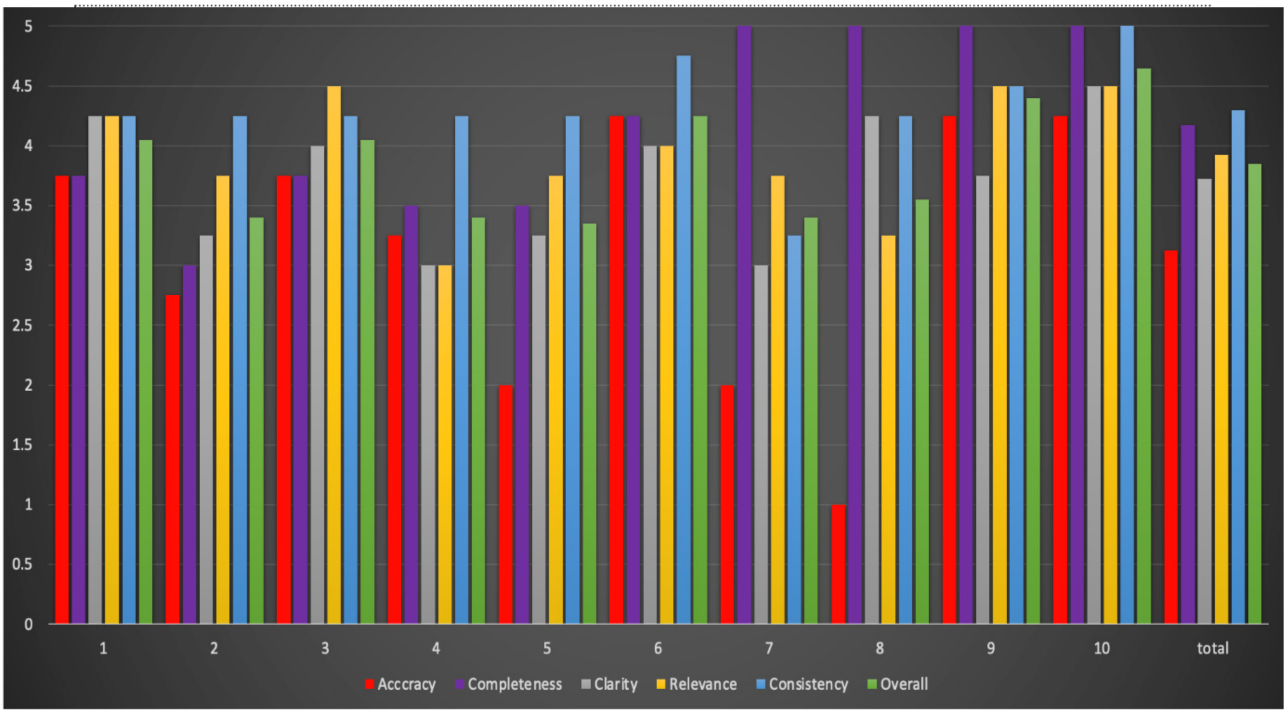
Figure 2 Parameters used to compare evaluation criteria across five main aspects based on the Global Quality Score (GQS): Accuracy, Completeness, Clarity, Relevance, and Consistency.

ตารางที่ 2 *ค่าเฉลี่ย Global Quality Score (GQS) แบ่งตามชนิดของแชทบอท*

Table 2 Average Global Quality Score (GQS) by Type of Chatbot

	Accuracy (Means±SD)	Completeness (Means±SD)	Clarity (Means±SD)	Relevance (Means±SD)	Consistency (Means±SD)
ChatGPT (GPT-4o)	2.80±1.23	4.10±1.20	3.50±0.97	3.80±0.63	4.20±0.79
Google Gemini (Gemini 2.0 Flash)	3.90±1.66	4.70±0.95	4.10±1.10	4.40±0.84	4.50±0.71
Anthropic Claude (Claude 3.5 Sonnet)	2.40±1.07	3.90±1.10	2.80±0.79	3.30±0.67	3.80±0.92
Microsoft Copilot (standard Copilot with o1 reasoning)	3.40±1.58	4.00±1.05	4.50±0.85	4.20±0.92	4.70±0.48
Total	3.13±1.47	4.18±1.08	3.73±1.11	3.93±0.86	4.30±0.79
P-value†	0.13	0.33	0.002*	0.03*	0.07

† เปรียบเทียบ one-way ANOVA ระหว่าง AICB ทั้ง 4 โปรแกรม



รูปที่ 3 *การประเมินแยกตามคำถามทั้ง 10 ข้อ เมื่อคิดค่าเฉลี่ยคะแนนของทุกโปรแกรมแชทบอทแยกตามคำถาม*

Figure 3 *Evaluation results by individual question across all 10 questions, showing the average score of each chatbot program per question*

บทวิจารณ์

จากผลการศึกษานี้สามารถสะท้อนให้เห็นถึงศักยภาพและข้อจำกัดของการใช้ภาษาไทยในการป้อนข้อมูลสู่ AICB ชนิดต่างๆ 4 โปรแกรมที่ขับเคลื่อนด้วย LLM ในการให้ข้อมูลทางทันตกรรมเฉพาะทางเกี่ยวกับ OGS โดยการประเมินคุณภาพคำตอบของ AICB ทั้ง 4 โปรแกรม ได้แก่ GPT-4o, Gemini 2.0 Flash, Claude 3.5 Sonnet และ Standard Copilot with o1 reasoning ผ่านเกณฑ์ GQS (ใน 5 ด้าน คือ ความถูกต้อง, ความครบถ้วน, ความชัดเจน, ความเกี่ยวข้อง และความสม่ำเสมอ) พบว่า AICB แต่ละตัวให้ผลลัพธ์ที่แตกต่างกัน โดยเฉพาะเมื่อพิจารณาค่าเฉลี่ยแต่ละด้าน ค่าเฉลี่ยของความถูกต้องโดยรวมอยู่ที่ 3.13±1.47 ซึ่งถือว่าอยู่ในระดับปานกลาง โดย Gemini 2.0 Flash ให้คะแนนสูงที่สุดที่ 3.90±1.66 ขณะที่ Claude 3.5 Sonnet มีคะแนนต่ำที่สุดที่ 2.40±1.07 แสดงถึงความแตกต่าง

ในการประมวลผลเนื้อหาทางคลินิก ในด้านความสมบูรณ์ ซึ่งมีค่าเฉลี่ยรวม 4.18 ± 1.08 นับว่าอยู่ในระดับดี Gemini 2.0 Flash ยังคงเป็นโปรแกรมที่ให้คะแนนสูงสุด (4.70 ± 0.95) ในขณะที่ Claude 3.5 Sonnet ให้ค่าต่ำสุดที่ 3.90 ± 1.10 ด้าน ความชัดเจนมีค่าเฉลี่ยรวมอยู่ที่ 3.73 ± 1.11 โดย Standard Copilot with o1 reasoning ให้คะแนนสูงสุดถึง 4.50 ± 0.85 ขณะที่ Claude 3.5 Sonnet ได้ต่ำสุดที่ 2.80 ± 0.79 สะท้อนให้เห็นถึงความแตกต่างในการเลือกใช้ภาษาที่เข้าใจง่ายและเหมาะสมกับผู้ป่วย ในด้านความเกี่ยวข้องค่ารวมเฉลี่ยอยู่ที่ 3.93 ± 0.86 โดย Gemini 2.0 Flash ได้คะแนนสูงสุดที่ 4.40 ± 0.84 ขณะที่ Claude 3.5 Sonnet อยู่ที่ 3.30 ± 0.67 ซึ่งจากการประเมินทางสถิติพบว่า 2 ด้านคือความชัดเจนและความเกี่ยวข้องมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ และสุดท้ายในด้าน ความสม่ำเสมอค่าเฉลี่ยรวมอยู่ที่ 4.30 ± 0.79 ซึ่งเป็นด้านที่มีคะแนนเฉลี่ยรวมสูงที่สุด โดย Standard Copilot with o1reasoning ได้คะแนนสูงสุดที่ 4.70 ± 0.48 ขณะที่ Claude 3.5 Sonnet ได้คะแนนต่ำสุดที่ 3.80 ± 0.92 ผลลัพธ์เหล่านี้ชี้ให้เห็นว่าแม้ AICB ทุกตัวจะสามารถตอบคำถามที่ครอบคลุมและมีความสม่ำเสมอในระดับหนึ่ง แต่ยังมีความแปรปรวนด้านความถูกต้องซึ่งอาจทำให้ผู้ป่วยได้รับข้อมูลที่ผิดพลาดจนนำไปสู่ผลข้างเคียงที่ไม่คาดคิดได้ เฉกเช่นเดียวกับความชัดเจนและความเกี่ยวข้องซึ่งถือเป็นองค์ประกอบสำคัญว่าของผู้ป่วยว่าจะเข้าใจชุดข้อมูลนั้นหรือไม่ก็ยังพบความแตกต่างของค่าเฉลี่ยของแต่ละโปรแกรมอย่างมีนัยสำคัญ เมื่อวิเคราะห์ลึกลงไปในระดับรายคำถามยังพบความแตกต่างในคุณภาพของคำตอบอีกระดับหนึ่ง เช่น คำตอบของคำถามข้อ 6 และข้อ 10 ได้คะแนนความถูกต้องสูงสุดที่ 4.25 คะแนน แสดงถึงความสามารถของ AICB ในการตอบคำถามเชิงคลินิกที่มีขอบเขตชัดเจนและเป็นที่ยอมรับในวงการแพทย์ ขณะที่คำถามข้อ 8 (สอนการดูแลความสะอาดแผลในช่องปากหลัง OGS) ได้คะแนนความถูกต้องต่ำสุดเพียง 1.00 คะแนน สาเหตุหลักๆ ที่คะแนนความถูกต้องต่ำในข้อนี้คือการแนะนำให้หลีกเลี่ยงการแปรงฟันหลังผ่าตัดตั้งแต่ 24 ชั่วโมงจนถึง 2 สัปดาห์ ซึ่งอาจเกิดจากความซับซ้อนของการดูแลหลังผ่าตัดที่ต้องการคำแนะนำเฉพาะบุคคล รวมถึงการขาดข้อมูลที่ถูกต้องแหล่งข้อมูลที่เป็นภาษาไทยที่ใช้เป็นแหล่งข้อมูลสำหรับ AICB ซึ่งอาจนำมาจากการเขียนบทความโดยผู้ที่เคยทำการผ่าตัดมาก่อนหน้าเอง และในเว็บไซต์ของสถานพยาบาลซึ่งอาจไม่ได้ถูกเขียนโดยผู้เชี่ยวชาญเอง[21,22] การวิจัยนี้จึงเน้นย้ำถึงความจำเป็นในการประเมินคุณภาพของคำตอบอย่างรอบด้าน และแสดงให้เห็นถึงความขาดแคลนของข้อมูลเกี่ยวกับการดูแลตนเองหลังผ่าตัดในโมเดลภาษาไทย ชี้ให้เห็นหน้าที่ของผู้เชี่ยวชาญเฉพาะทางโดยเฉพาะอย่างยิ่งศัลยแพทย์ช่องปากและแม็กซิลโลเฟเชียล หรือทันตแพทย์จัดฟันในการการเขียนบทความภาษาไทย เพื่อสร้างข้อมูลที่ถูกต้องน่าเชื่อถือเพิ่มเข้าไปใน LLM ของ AICB ต่างๆ

งานวิจัยนี้ได้คัดเลือกAICB ที่เป็นที่นิยมใช้ทั่วไปจำนวน 4 AICB โปรแกรมดังกล่าวถูกเลือกเนื่องจากข้อมูลการใช้งานและปริมาณการเข้าถึงที่เปิดเผยต่อสาธารณะล่าสุดแสดงให้เห็นว่าโปรแกรมเหล่านี้รวมกันครองส่วนแบ่งตลาดของ AICB ด้านการสนทนาเป็นส่วนใหญ่ และมีการใช้งานอย่างแพร่หลายในกลุ่มผู้ใช้ทั่วไปที่ไม่ใช่ผู้เชี่ยวชาญด้านเทคโนโลยี ทั้ง 4 AICB นี้ ทำให้เราสามารถครอบคลุมเครื่องมือที่ได้รับความนิยมสูงสุดในหมู่ผู้บริโภค16 แม้ว่าจะมีแชทบอทอื่น ๆ อีก เช่น Perplexity (Perplexity, Perplexity AI, San Francisco, CA) หรือ Deepseek (DeepSeek-V2, DeepSeek AI, Shanghai, China) อย่างไรก็ตามควรตระหนักว่าส่วนแบ่งตลาดของระบบเหล่านี้มีการเปลี่ยนแปลงอยู่ตลอดเวลา ดังนั้นงานวิจัยนี้จึงสะท้อนให้เห็นถึงภาพรวม ณ ช่วงเวลากลางปี พ.ศ.2568 เท่านั้น เมื่อเปรียบเทียบกับการศึกษาในภาษาอังกฤษที่คล้ายกัน ในวรรณกรรมส่วนใหญ่ที่เกี่ยวข้องกับ AICB ทางการแพทย์ มักมุ่งเน้นไปที่การศึกษาการทำงานของ ChatGPT โดยหนึ่งในการศึกษาครั้งแรกๆ ChatGPT ได้ทำการวิเคราะห์เนื้อหาของคำตอบที่สร้างโดย AI เกี่ยวกับการจัดฟันด้วยเครื่องมือใส (orthodontic clear aligners) โดยพบว่าความถูกต้องโดยรวมของคำตอบจากเวอร์ชันแรกเริ่มของ ChatGPT ยังไม่เพียงพอ และไม่มีการอ้างอิงแหล่งข้อมูลวิชาการที่ชัดเจน พร้อมเน้นย้ำข้อจำกัดของAICBในการให้ข้อมูลที่เป็นปัจจุบันและถูกต้อง[23] Duran และคณะ[24] รายงานว่า ChatGPT สามารถให้ข้อมูลที่มีความน่าเชื่อถือและคุณภาพในระดับสูงเกี่ยวกับภาวะปากแหว่งเพดานโหว่ แต่มีความยากในการอ่านเข้าใจและข้อมูลที่ได้รับควรถูกตรวจสอบความถูกต้องโดยผู้เชี่ยวชาญทางการแพทย์ ขณะที่ Kılıç และคณะ[21] ได้ศึกษาความน่าเชื่อถือและความสามารถในการอ่านของคำตอบต่อคำถามทางทันตกรรมจัดฟันโดยใช้ ChatGPT เปรียบเทียบระหว่างรุ่นใหม่ และรุ่นก่อนหน้าพบว่าความน่าเชื่อถือของคำตอบอยู่ในระดับปานกลางและสรุปว่า ChatGPT ยังไม่อาจถือว่ามีความถูกต้องทางวิทยาศาสตร์ได้ เนื่องจากไม่มีการอ้างอิงแหล่งข้อมูล Demirsoy และคณะ[22] ประเมินความน่าเชื่อถือของข้อมูลที่ผลิตโดย ChatGPT-4 โดยพิจารณาจากความถูกต้องและความเกี่ยวข้องของคำตอบ โดยให้ทันตแพทย์จัดฟันนักศึกษาทันตแพทย์ และผู้ป่วยเป็นผู้ประเมิน ผลการศึกษาสรุปว่า ChatGPT มีศักยภาพที่สำคัญในการให้ข้อมูลและให้ความรู้แก่ผู้ป่วยหากได้รับการพัฒนาและปรับปรุงแก้ไขอย่างเหมาะสม สำหรับการศึกษานี้ พบว่า ChatGPT-4o ซึ่งเป็นเวอร์ชันปัจจุบัน สามารถให้คำตอบเกี่ยวกับการถอนฟันเพื่อการจัดฟันในระดับที่มีคุณภาพและความถูกต้องดี ซึ่งแสดงให้เห็นว่าระบบภาษาที่พัฒนาขึ้นของ LLM อาจสามารถให้คำตอบที่แม่นยำและมีคุณภาพยิ่งขึ้น การศึกษาหนึ่ง

ประเมินความสามารถด้านความชัดเจนด้านการอ่าน(readability) ของข้อมูลที่สร้างโดย ChatGPT-3.5, ChatGPT-4, Gemini และ Copilot เกี่ยวกับเครื่องมือจัดฟันแบบใส โดยใช้ดัชนี Flesch Reading Ease Score (FRES) ผลการศึกษาพบว่าคำตอบจาก AICB ที่ใช้ AI ทั้งหมดมีระดับความยากในการอ่านค่อนข้างสูง ซึ่งสอดคล้องกับผลการประเมินด้านความสามารถในการอ่านของการศึกษานี้ ทั้งนี้ผู้วิจัยเสนอว่า ข้อความควรถูกปรับให้ง่ายขึ้นโดยการลดจำนวนประโยคที่ยาวและลดการใช้คำศัพท์ที่ซับซ้อน เพื่อส่งเสริมความเข้าใจและความสามารถในการเข้าถึงข้อมูลของผู้รับสารได้ดียิ่งขึ้น[25] สิ่งหนึ่งที่น่าสังเกตในการศึกษานี้คือการเลือกปริมาณคำถามที่จำกัด การประเมินแบบสอบถามที่จำกัดเพียงผู้เชี่ยวชาญเท่านั้นและใช้เครื่องมือในการวัดผลเพียงชนิดเดียวคือ GQS การหาเครื่องมืออื่นๆ เช่น DISCERN[26] หรือ mDISCERN[27] อาจถูกนำมาใช้ประกอบในการศึกษาต่อๆ ไป ขณะเดียวกันการเพิ่มกลุ่มบุคคลทั่วไป เข้ามาเป็นผู้ประเมินอีกกลุ่มหนึ่ง อาจช่วยเสริมความแข็งแกร่งให้กับผลการวิจัยได้ ยิ่งไปกว่านั้น ในการประเมินด้านความชัดเจนการมีส่วนร่วมจากบุคคลทั่วไปจะเป็นประโยชน์อย่างยิ่ง เนื่องจากพวกเขาอาจสามารถระบุศัพท์เทคนิคหรือภาษาเฉพาะทางที่ผู้เชี่ยวชาญอาจมองข้ามไป[22,25] แต่ถึงอย่างไรการที่ผู้วิจัยให้ผลสอดคล้องกันน่าจะเกิดจากการพยายามมองในมุมมองของบุคคลทั่วไปในหัวข้อนี้

ปัญหาหนึ่งที่พบได้บ่อยในการศึกษาที่ประเมินความถูกต้องของคำตอบจาก AICB คือการให้ข้อมูลที่คลาดเคลื่อนหรือไม่ถูกต้องตามข้อเท็จจริงซึ่งเรียกว่าภาวะ Hallucination ปรากฏการณ์นี้ไม่ใช่การจงใจให้ข้อมูลที่เท็จ แต่เป็นผลจากกลไกการทำงานที่เน้นการทำนายลำดับคำถัดไปที่มีความเป็นไปได้ทางสถิติสูงสุด (Probabilistic Next-Token Prediction) มากกว่าการยึดโยงกับข้อเท็จจริง (Grounding) ปรากฏการณ์นี้ในบริบททางการแพทย์ถือว่ามีความเสี่ยงสูงมากเพราะคำตอบที่ไม่ถูกต้องเพียงเล็กน้อย เช่น ข้อมูลยาผิด หรือแนวทางการรักษาที่ไม่เป็นไปตามมาตรฐาน อาจนำไปสู่ผลลัพธ์ที่เป็นอันตรายถึงชีวิตได้ โดยคำตอบในลักษณะนี้มักมีรูปแบบประโยคที่ถูกต้องทั้งในเชิงไวยากรณ์และความหมาย แต่เนื้อหากลับผิดจากข้อเท็จจริงหรือขาดสาระสำคัญ[28,29] สาเหตุที่อาจนำไปสู่การเกิด hallucination มีได้หลายประการ เช่น ข้อจำกัดของกระบวนการสร้างข้อความของ LLM การขาดการอัปเดตข้อมูลแบบเรียลไทม์ การไม่มีการเข้าถึงฐานข้อมูลทางวิชาชีพ และความซับซ้อนของบริบททางคลินิกที่ต้องใช้ความเข้าใจในหลายปัจจัยร่วมกัน รวมไปถึงข้อมูลวิชาการภาษาไทยคุณภาพสูงโดยผู้เชี่ยวชาญ ซึ่งเป็นกลไกสำคัญในการ Grounding ความรู้ของโมเดลเข้ากับข้อเท็จจริงที่ตรวจสอบได้ ในการศึกษาของ Alkuraya และคณะ[30] พบว่า AICB ChatGPT, Claude และ Bard สามารถระบุรูปแบบการถ่ายทอดทางพันธุกรรมของโรคได้อย่างถูกต้อง แต่กลับคำนวณความน่าจะเป็นของการมีบุตรที่มีสุขภาพดีได้ไม่ถูกต้อง ซึ่งอาจเกิดจากข้อจำกัดของ AICB ในการจัดการกับปัญหาทางการแพทย์ที่มีความซับซ้อนและต้องอาศัยตัวแปรหลายปัจจัยในการวิเคราะห์อย่างถูกต้อง ในการศึกษานี้จะมีคำตอบหนึ่งของ Claude 3.5 Sonnet ในหัวข้อการเลื่อนคาง “จำเป็นต้องทำพร้อมกับการเลื่อนขากรรไกรหรือไม่ เพราะเหตุใด” โดย Claude 3.5 Sonnet ตอบในทางตรงกันข้ามว่าการเลื่อนคางสามารถทำการเลื่อนขากรรไกรไปพร้อมกันได้ หมายถึงการทำหัตถการเลื่อนคางเป็นหัตถการหลัก และ OGS เป็นหัตถการเสริม (adjunctive treatment) ด้วยเหตุนี้การศึกษานี้จึงไม่เพียงแต่ชี้ให้เห็นถึงความแตกต่างระหว่าง AICB แต่ละตัวแต่ยังสะท้อนถึงข้อจำกัดของเทคโนโลยี LLM ในภาษาท้องถิ่น[31] ซึ่งรายงานฉบับนี้มุ่งให้ข้อมูลด้านสุขภาพในบริบทของภาษาไทย ซึ่งยังขาดงานวิชาการภาษาไทยที่ LLM สามารถเข้าถึงเพื่อนำไปสังเคราะห์เป็นบทความใหม่ได้ การยกระดับคุณภาพการตอบสนองภาษาไทยของระบบ AICB จึงมิอาจพึ่งพาเพียงการถ่ายโอนความรู้ข้ามภาษา (Cross-Lingual Transfer) จากแหล่งข้อมูลต่างประเทศได้เท่านั้น หากแต่จำเป็นต้องลงทุนสร้างข้อมูลต้นน้ำที่เป็นบทความวิชาการภาษาไทยคุณภาพสูง เช่น บทความจากวารสารทางการแพทย์ที่ผ่านการตรวจสอบโดยผู้เชี่ยวชาญ มีความถูกต้องทางวิชาการ (Academic Fidelity) และความเหมาะสมเชิงวัฒนธรรม (Cultural Appropriateness) เพื่อเป็นเสาหลักในการแก้ไขปัญหาความเหลื่อมล้ำทางข้อมูล (Data Inequity) และลดความเสี่ยงของการเกิดอาการหลอน[32-35] อันจะนำไปสู่การตัดสินใจของผู้ป่วยที่ปลอดภัยและเกิดประโยชน์สูงสุดอย่างแท้จริง ทั้งนี้คุณภาพของคำตอบจาก AICB มีลักษณะเป็นพลวัต เนื่องจากระบบมีการพัฒนาข้อมูลและโมเดลอย่างต่อเนื่อง ทำให้ผลการวิจัยในช่วงเวลาหนึ่งอาจไม่สะท้อนคุณภาพของโปรแกรมในอนาคตได้อย่างแท้จริง ในแง่มุมนี้ก็เป็นโอกาสที่ฝ่ายกำหนดนโยบายควรตระหนักว่าในยุคแห่งปัญญาประดิษฐ์การลงทุนในข้อมูลทางการแพทย์ภาษาไทยคุณภาพสูง คือการลงทุนในความปลอดภัยและคุณภาพชีวิตของผู้ใช้ภาษาไทยในอนาคต

สรุปผลการศึกษา

การศึกษานี้ประเมินคุณภาพคำตอบจากระบบ AICB ภาษาไทยที่ขับเคลื่อนด้วย LLM ในการให้ข้อมูลเกี่ยวกับ OGS พบว่าแต่ละ AICB มีจุดเด่นและข้อจำกัดแตกต่างกันไป โดย Google Gemini 2.0 Flash ได้คะแนนเฉลี่ยสูงสุดในหลายด้าน ขณะที่ Microsoft Copilot โดดเด่นในเรื่องความชัดเจนและความสม่ำเสมอ ส่วน ChatGPT-4o สามารถสื่อสารได้เข้าใจง่าย เหมาะสำหรับผู้ไม่มีพื้นฐานทางการแพทย์ ในขณะที่ Claude 3.5 Sonnet มีคะแนนต่ำสุดในหลายด้าน สะท้อนข้อจำกัดในการประมวลผลข้อมูลทางการแพทย์ในภาษาไทย ผลการศึกษาพบว่าการพัฒนาเนื้อหาวิชาการภาษาไทยที่มีคุณภาพและเข้าถึงได้โดย LLM มีความสำคัญอย่างยิ่งเพื่อให้ระบบสามารถสังเคราะห์ข้อมูลที่ถูกต้องแม่นยำและเหมาะสม

กับบริบททางวัฒนธรรมของไทย ทั้งนี้ AICB ภาษาไทยยังคงมีศักยภาพในการให้ข้อมูลเบื้องต้นแก่ผู้ป่วยเกี่ยวกับ OGS แต่ควรใช้เป็นเครื่องมือเสริมด้วยความระมัดระวัง พร้อมแนะนำให้ผู้ป่วยปรึกษาผู้เชี่ยวชาญทางสุขภาพเพื่อประกอบการตัดสินใจอย่างถูกต้องและปลอดภัย นอกจากนี้ผู้วิจัยขอเสนอให้พัฒนาระบบแนวปฏิบัติ (Guideline) สำหรับบุคลากรทางการแพทย์เพื่อแนะนำผู้ป่วยในการใช้ AICB ค้นหาข้อมูลสุขภาพอย่างปลอดภัยและมีประสิทธิภาพเพื่อส่งเสริมการใช้งานเทคโนโลยีนี้ในอย่างเหมาะสมและลดความเสี่ยงจากการตีความข้อมูลผิดพลาด

กิตติกรรมประกาศ

ขอบคุณ ศ.เกียรติคุณ ทพ.ณัฐเมศร์ วงศ์สิริฉัตร, รศ.ดร.รัชชพิน ศรีสัจจะลักษณ์ และคณาจารย์คณะทันแพทยศาสตร์ มหาวิทยาลัยกรุงเทพธนบุรี

เอกสารอ้างอิง

1. Huh JK. Orthognathic surgery of temporomandibular disorders. ***J Korean Assoc Oral Maxillofac Surg*** 2021;47(2):63-4.
2. Zammit D, Ettinger RE, Sanati-Mehrizy P, Susarla SM. Current trends in orthognathic surgery. ***Medicina (Kaunas)*** 2023;59(12):2100.
3. Pithon MM, dos Santos ES. Information available on the internet about pain after orthognathic surgery: a careful review. ***Dental Press J Orthod*** 2014;19(6):86-92.
4. Engelmann J, Fischer C, Nkenke E. Quality assessment of patient information on orthognathic surgery on the internet. ***J Craniomaxillofac Surg*** 2020;48(7):661-5.
5. Hegarty E, Campbell C, Grammatopoulos E, DiBiase AT, Sherriff M, Cobourne MT. YouTube™ as an information resource for orthognathic surgery. ***J Orthod*** 2017;44(2):90-6.
6. Bhamrah G, Ahmad S, NiMhurchadha S. Internet discussion forums, an information and support resource for orthognathic patients. ***Am J Orthod Dentofacial Orthop*** 2015;147(1):89-96.
7. Larsen MK, Thygesen TH. Orthognathic surgery: outcome in a Facebook group. ***J Craniofac Surg*** 2016;27(2):350-5.
8. Alsuhaym O, Aldawas I, Maki F, Alamro M, Alshehri K, Alharthi Y. Does social media affect a patient's decision to undergo orthognathic surgery? ***Int J Environ Res Public Health*** 2023;20(12):6103.
9. Chowdhury N, Khalid A, Turin TC. Understanding misinformation infodemic during public health emergencies due to large-scale disease outbreaks: a rapid review. ***Z Gesundh Wiss*** 2023;31(4):553-73.
10. Paez A. Gray literature: An important resource in systematic reviews. ***J Evid Based Med*** 2017;10(3):233-40.
11. Buyuk SK, Imamoglu T. Instagram as a social media tool about orthognathic surgery. ***Health Promot Perspect*** 2019;9(4):319-22.
12. Strunga M, Urban R, Surovková J, Thurzo A. artificial intelligence systems assisting in the assessment of the course and retention of orthodontic treatment. ***Healthcare (Basel)*** 2023;11(5):683.
13. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. ***J Esthet Restor Dent*** 2023;35(7):1098-102.
14. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, *et al.* A systematic review of large language model (LLM) evaluations in clinical medicine. ***BMC Med Inform Decis Mak*** 2025;25(1):117.
15. Thurzo A, Urbanová W, Novák B, Czako L, Siebert T, Stano P, *et al.* Where is the artificial intelligence applied in dentistry? systematic review and literature analysis. ***Healthcare (Basel)*** 2022;10(7):1269.
16. StatCounter Global Stats. AI Chatbot Market Share [Internet]. Dublin: StatCounter; 2025 [cited 2025 Oct 19]. Available from: https://gs.statcounter.com/ai-chatbot-market-share#quarterly-202503-202503-bar
17. Hochmair HH, Juhász L, Kemp T. Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks. Transactions in GIS. 2024 Nov;28(7):2219-31.
18. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. Adv Physiol Educ. 2025 Jun 1;49(2):430-437. doi: 10.1152/advan.00093.2024. Epub 2025 Jan 17. PMID: 39824512.
19. Feinberg I, Ogrodnick M, Bernhardt J. COVID-19 Vaccine Videos: Health Literacy Considerations. Health Lit Res Pract. 2023 Jun; 7(2):e111-e118. doi: 10.3928/24748307-20230523-02. Epub 2023 Jun 1. PMID: 37306321; PMCID: PMC10256273.
20. Hyland ME, Sodergren SC. Development of a new type of global quality of life scale, and comparison of performance and preference for 12 global scales. ***Qual Life Res*** 1996;5(5):469-80.
21. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. ***Am J Orthod Dentofacial Orthop*** 2024; 165(5):546-55.
22. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? ***Angle Orthod*** 2024;94(6):602-7.
23. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. ***Angle Orthod*** 2024;94(3):263-72.
24. Duran GS, Yurdakurban E, Topsakal KG. The quality of CLP-related Information for patients provided by ChatGPT. ***Cleft Palate Craniofac J*** 2025;62(4):588-95.
25. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? ***BMC Med Inform Decis Mak*** 2024;24(1):211.
26. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol. 2023 Oct 1;9(10):1437-1440. doi: 10.1001/jamaoncol.2023.2947. PMID: 37615960; PMCID: PMC10450581.

27. Senbaykal Yigit E, Taskirdi I, Haci IA, Tuncel T. Artificial intelligence performance in pediatric asthma. J Asthma. 2025 Aug 1:1-7. doi: 10.1080/02770903.2025.2531500. Epub ahead of print. PMID: 40643318.
28. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. ***Cureus*** 2023;15(4):e37432.
29. Graf EM, McKinney JA, Dye AB, Lin L, Sanchez-Ramos L. Exploring the limits of artificial intelligence for referencing scientific articles. ***Am J Perinatol*** 2024;41(15):2072-81.
30. Alkuraya IF. Is artificial intelligence getting too much credit in medical genetics? ***Am J Med Genet C Semin Med Genet*** 2023;193(3):e32062.
31. Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. ***J Stomatol Oral Maxillofac Surg*** 2024;125(5):101757.
32. Google DeepMind. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context(Technical Report). Google DeepMind. Retrieved from https://www.reportpdf.com/.
33. Lewis, P., Yih, W., Ghazvininejad, M., Mohr, S., & Goyal, A. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
34. Anthropic. (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku - Model Card. Anthropic.URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
35. Anthropic. (2024). Multilingual Support - Claude Documentation. URL: https://docs.claude.com/en/docs/build-with-claude/multilingual-support