

บทบาทของเทคนิคการจัดกลุ่ม  
และสถาปัตยกรรม MIXTURE OF EXPERTS เบื้องหลังความสำเร็จของ DEEPSEEK  
THE ROLE OF CLUSTERING TECHNIQUES AND MIXTURE OF EXPERTS  
ARCHITECTURE BEHIND THE SUCCESS OF DEEPSEEK

สุริยะ พุ่มเฉลิม<sup>1\*</sup>, นพคุณ นันทเสนีย์<sup>2</sup> และ อุษณี ยี่สุนแก้ว<sup>3</sup>

Suriya Pumchalerm<sup>1\*</sup>, Nopakhun Nanthaseneey<sup>2</sup> and Utsanee Yeeseenkaew<sup>3</sup>

คณะเทคโนโลยีดิจิทัลและนวัตกรรม มหาวิทยาลัยเซาท์อีสต์ Bangkok<sup>1,2,3</sup>

Faculty of Digital Technology and Innovation, Southeast Bangkok University<sup>1,2,3</sup>

Author email: pumchalerm@hotmail.com<sup>1\*</sup>, 6711463100006@sbu.southeast.ac.th<sup>2</sup>,

utsanee@sbu.southeast.ac.th<sup>3</sup>

Received: February 24, 2026

Revised: April 8, 2026

Accepted: April 28, 2026

บทคัดย่อ

บทความนี้เป็นบทความวิชาการเชิงทบทวนและสังเคราะห์แนวคิด โดยมีวัตถุประสงค์เพื่อนำเสนอแนวทางการพัฒนาปัญญาประดิษฐ์ ซึ่งในปัจจุบันกำลังเผชิญความท้าทายของข้อจำกัดในด้านความซับซ้อนเชิงคำนวณ การใช้ทรัพยากรจำนวนมาก และการขยายขนาดของระบบ ดังนั้นผู้ออกแบบระบบปัญญาประดิษฐ์จึงมุ่งพัฒนาแนวคิดและสถาปัตยกรรมที่สามารถรักษาประสิทธิภาพการประมวลผลในขณะที่ลดต้นทุนทรัพยากร และแนวทางที่ได้รับความนิยม คือ เทคนิคการจัดกลุ่ม ที่เป็นการเรียนรู้ของเครื่องจักรรูปแบบหนึ่ง และสถาปัตยกรรม Mixture of Experts (MoE) ซึ่งจะทำหน้าที่วิเคราะห์ความเชื่อมโยงเชิงโครงสร้างและเชิงคำนวณ โดยเทคนิคการจัดกลุ่มจะมุ่งค้นหารูปแบบและโครงสร้างแฝงภายในข้อมูล โดยอาศัยการแทนความหมายเชิงลึกเพื่อจัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกัน กลไกดังกล่าวช่วยลดความซับซ้อนของข้อมูลอินพุตและเพิ่มประสิทธิภาพในการประมวลผล ขณะเดียวกัน สถาปัตยกรรม MoE เป็นแนวคิดการคำนวณเชิงเงื่อนไขที่แบ่งแบบจำลองออกเป็นผู้เชี่ยวชาญย่อยหลายส่วน และใช้กลไกเลือกเปิดใช้งานเฉพาะส่วนที่เหมาะสมกับข้อมูล ส่งผลให้เกิดการใช้ทรัพยากรอย่างมีประสิทธิภาพและรองรับการขยายขนาดของระบบได้ดียิ่งขึ้น ด้วยเหตุนี้การผสมผสานเทคนิคการจัดกลุ่มเข้ากับสถาปัตยกรรม MoE จึงก่อให้เกิดโครงสร้างการคำนวณที่สามารถลดภาระเชิงคำนวณ เพิ่มความยืดหยุ่น และรักษาสมรรถนะของแบบจำลองในระดับสูง แนวคิดดังกล่าวสะท้อนให้เห็นทิศทางสำคัญของการพัฒนาระบบปัญญาประดิษฐ์ “DeepSeek” ที่มุ่งเน้นทั้งประสิทธิภาพ ความสามารถในการปรับขนาด และความยั่งยืนในการใช้งานในบริบทที่หลากหลาย

คำสำคัญ: เทคนิคการจัดกลุ่ม, Mixture of Experts, แบบจำลองภาษาขนาดใหญ่, ดีพีซี

## Abstract

This article is an academic review and synthesis of concepts, with the objective of presenting guidelines for the development of artificial intelligence (AI), which is currently confronted with significant challenges, particularly constraints related to computational complexity, intensive resource consumption, and system scalability. Consequently, AI system designers have increasingly focused on developing conceptual frameworks and architectural paradigms that can maintain high computational performance while reducing resource costs. Among the approaches that have attracted considerable attention are clustering techniques, one form of machine learning, and the Mixture of Experts (MoE) architecture, which functions to analyze structural and computational interrelationships within models. Clustering techniques aim to discover latent patterns and intrinsic structures within data by leveraging deep representations to group instances with similar characteristics. This mechanism helps reduce the complexity of input data and enhances processing efficiency. In parallel, the MoE architecture embodies a conditional computation paradigm in which a model is divided into multiple specialized sub-networks, or “experts,” and a routing mechanism selectively activates only those components most relevant to a given input. Such an approach enables more efficient resource utilization and improves the scalability of large-scale systems. The integration of clustering techniques with the MoE architecture therefore yields a computational framework capable of reducing computational burden, increasing flexibility, and sustaining high model performance. This conceptual integration reflects a key direction in the development of the AI system “DeepSeek,” which emphasizes efficiency, scalability, and sustainable deployment across diverse application contexts.

**Keywords:** Clustering, Mixture of Experts, Large Language Model, DeepSeek

## บทนำ

บทความนี้เป็นบทความวิชาการเชิงทบทวนและสังเคราะห์แนวคิด โดยมีวัตถุประสงค์เพื่อนำเสนอแนวทางการพัฒนาระบบปัญญาประดิษฐ์ โดยใช้เทคนิคการจัดกลุ่ม (Clustering) และสถาปัตยกรรม Mixture of Experts (MoE) เป็นแนวคิดสำคัญที่อยู่เบื้องหลังความสำเร็จของ DeepSeek โดย Clustering ช่วยจัดระเบียบข้อมูลออกเป็นกลุ่มที่มีลักษณะคล้ายกัน ทำให้โมเดลสามารถเรียนรู้ได้อย่างมีประสิทธิภาพและลดความซับซ้อนของปัญหา ในขณะที่ MoE เป็นการออกแบบโมเดลให้มี “ผู้เชี่ยวชาญหลายตัว” ที่มีความสามารถเฉพาะด้าน เพื่อเลือกใช้เฉพาะส่วนที่เหมาะสมกับข้อมูลในแต่ละครั้ง แนวทางนี้ช่วยเพิ่มประสิทธิภาพในการประมวลผล ลดการใช้ทรัพยากร และเพิ่มความแม่นยำของผลลัพธ์ นอกจากนี้ การทำงานร่วมกันของ Clustering และ

MoE ยังช่วยให้โมเดลสามารถจัดการข้อมูลขนาดใหญ่และหลากหลายได้ดีขึ้น ส่งผลให้ DeepSeek มีความสามารถทั้งด้านความลึกและความยืดหยุ่น และสามารถแข่งขันกับเทคโนโลยีจากองค์กรชั้นนำได้อย่างมีประสิทธิภาพ

การเรียนรู้ของเครื่องจักรและการเรียนรู้เชิงลึกได้กลายเป็นรากฐานสำคัญของระบบสารสนเทศและระบบปัญญาประดิษฐ์ในยุคดิจิทัล ตั้งแต่ระบบค้นคืนข้อมูล ระบบแนะนำอัจฉริยะ ระบบรู้จำภาพและเสียง ไปจนถึงระบบประมวลผลภาษาธรรมชาติ การพัฒนาเทคโนโลยีเหล่านี้ต้องอาศัยความก้าวหน้าทั้งในด้านอัลกอริทึม โครงสร้างข้อมูล และสถาปัตยกรรมคอมพิวเตอร์ควบคู่กัน โดยสถาปัตยกรรมหรือโครงสร้างของปัญญาประดิษฐ์ ได้เปลี่ยนแปลงกระบวนการทัศน์ของการออกแบบอัลกอริทึมอย่างมีนัยสำคัญ โดยจะมาแทนที่โครงสร้างแบบวนซ้ำด้วยกลไกความสนใจ (Attention Mechanism) ซึ่งสามารถประมวลผลข้อมูลแบบขนานและเรียนรู้ความสัมพันธ์อย่างมีประสิทธิภาพ [1] โดยแนวคิดดังกล่าวไม่เพียงยกระดับสมรรถนะของระบบประมวลผลภาษาธรรมชาติ แต่ยังถูกนำไปประยุกต์ใช้ในสาขาอื่นของวิทยาการคอมพิวเตอร์อีกด้วย เช่น การมองเห็นด้วยคอมพิวเตอร์ (Computer Vision) และชีวสารสนเทศ (Bioinformatics) เป็นต้น

ดังนั้นความสำเร็จของการออกแบบอัลกอริทึมของปัญญาประดิษฐ์ จึงได้จุดประกายแนวโน้มการพัฒนาแบบจำลองที่อาศัยการเพิ่มจำนวนพารามิเตอร์และข้อมูลฝึกฝนเป็นกลยุทธ์หลัก ดังปรากฏให้เห็นได้จากงานวิจัยหลายฉบับที่แสดงให้เห็นความสัมพันธ์เชิงบวกระหว่างขนาดของแบบจำลองกับสมรรถนะในการทำงานเชิงภาษา [2] อย่างไรก็ตาม การขยายขนาดแบบจำลองดังกล่าวเริ่มสร้างภาระอย่างมีนัยสำคัญต่อระบบคอมพิวเตอร์และโครงสร้างพื้นฐานด้านเทคโนโลยีสารสนเทศ ที่ต้องการการใช้ทรัพยากรสูงขึ้นตามไปด้วย จากบริบทดังกล่าว จึงเกิดแนวคิดในการออกแบบอัลกอริทึมที่เน้นประสิทธิภาพเชิงโครงสร้างและการใช้ทรัพยากรอย่างเหมาะสม เมื่อเป็นเช่นนั้นกลุ่มนักวิทยาการคอมพิวเตอร์จึงให้ความสนใจกับ สถาปัตยกรรมแบบ Mixture of Experts มากขึ้น และถือได้ว่าเป็นหนึ่งในแนวทางที่ได้รับการยอมรับในวงกว้าง เนื่องจากสามารถเพิ่มขนาดแบบจำลองในเชิงพารามิเตอร์ได้โดยไม่เพิ่มภาระการคำนวณในสัดส่วนเดียวกัน [3]

### พัฒนาการของแบบจำลองภาษาขนาดใหญ่ (Large Language Model : LLM)

การพัฒนาระบบปัญญาประดิษฐ์นั้นทีมผู้พัฒนาระบบจะใช้ภาษาคอมพิวเตอร์หลายหลายภาษา และไม่ได้จำกัดด้วยภาษาเดียว ทั้งนี้ขึ้นอยู่กับวัตถุประสงค์ ประสิทธิภาพที่ต้องการ และสภาพแวดล้อมของระบบ แต่ทั้งนี้โดยส่วนใหญ่ของระบบปัญญาประดิษฐ์ที่สร้างขึ้นมาในปัจจุบันมักนิยมใช้ Python เป็นภาษาหลักในการพัฒนาระบบปัญญาประดิษฐ์ โดยเฉพาะระบบที่เน้นการทำงานแบบ Deep Learning และ Large Language Models เนื่องจากมีระบบนิเวศของเครื่องมือที่ครบถ้วน [4] โดย Large Language Model: LLM คือ แบบจำลองภาษาขนาดใหญ่ที่พัฒนาด้วยเทคนิค Deep Learning เพื่อให้คอมพิวเตอร์สามารถเข้าใจ สร้าง และโต้ตอบด้วยภาษามนุษย์ได้อย่างมีประสิทธิภาพและรองรับการทำงานภาษาของมนุษย์หลายภาษาได้ [2] ดังนั้นการพัฒนาระบบ Generative AI ในยุคปัจจุบันต้องอาศัย LLM เป็นกลไกหลักในการประมวลผลภาษาและความรู้เชิงข้อความ และนอกจากนี้ LLM จะทำหน้าที่เป็นแกนกลางของสถาปัตยกรรม ในขณะที่ยังประกอบอื่น ๆ จะช่วยสนับสนุนการจัดการข้อมูล การปรับขนาด และการประยุกต์ใช้งานจริง [5] การพัฒนาระบบปัญญาประดิษฐ์นั้นไม่เพียงขึ้นอยู่กับการพัฒนา

แบบจำลองภาษาขนาดใหญ่ แต่ต้องมีรากฐานมาจากการเรียนรู้เชิงสถิติและแบบจำลองความน่าจะเป็น โดยก่อนที่จะพัฒนาไปสู่การใช้เครือข่ายประสาทเทียมเชิงลึก ซึ่งถือได้ว่าเป็นเทคนิคหนึ่งที่สำคัญของระบบปัญญาประดิษฐ์ การเปลี่ยนผ่านจากแบบจำลองเชิงสถิติแบบดั้งเดิมไปสู่แบบจำลองเชิงลึก สะท้อนให้เห็นถึงบทบาทของการเพิ่มขีดความสามารถในการแทนความหมายของข้อมูล [6] ทั้งนี้แบบจำลองภาษายุคใหม่ เช่น Generative Pre-trained Transformer :GPT และ Bidirectional Encoder Representations from Transformers :BERT แสดงให้เห็นว่าการฝึกแบบจำลองบนข้อมูลขนาดใหญ่ร่วมกับสถาปัตยกรรมที่เหมาะสม สามารถสร้างองค์ความรู้เชิงภาษาที่นำไปประยุกต์ใช้ได้หลากหลาย อย่างไรก็ตาม แม้จะเห็นได้ว่าระบบปัญญาประดิษฐ์จะทำงานได้อย่างมีประสิทธิภาพ ตอบสนองการทำงานได้ตามความต้องการของผู้ใช้งาน แต่การพัฒนาในทิศทางดังกล่าวยังคงเผชิญข้อจำกัดด้านการใช้ทรัพยากรอยู่เช่นเดิม ดังนั้นจึงยังมีการค้นหาแนวทางการออกแบบที่มีประสิทธิภาพมากขึ้นในเชิงระบบ [5] ต่อไป

### เทคนิคการจัดกลุ่ม (Clustering)

การเรียนรู้ของเครื่องจักรมี 3 ประเภท คือ 1) Supervised Learning เป็นการเรียนรู้แบบมีผู้สอน 2) Unsupervised Learning เป็นการเรียนรู้แบบไม่มีผู้สอน และ 3) Reinforcement Learning เป็นการเรียนรู้ที่ให้ตัวแทนเรียนรู้ผ่านการโต้ตอบกับสภาพแวดล้อม [7] โดยเทคนิคการจัดกลุ่ม (Clustering) เป็นการเรียนรู้ของเครื่องจักรแบบไม่มีผู้สอน ซึ่งมีวัตถุประสงค์เพื่อแบ่งชุดข้อมูลออกเป็นกลุ่มย่อยโดยอาศัยความคล้ายคลึงหรือความสัมพันธ์ภายในข้อมูล โดยไม่ต้องอาศัยชุดข้อมูลฝึก [8] ซึ่งเทคนิคนี้มีบทบาทสำคัญในการทำความเข้าใจกับโครงสร้างของข้อมูลที่มีจำนวนมากๆ ได้เป็นอย่างดี เมื่อพิจารณาจากมุมมองเชิงทฤษฎี จะพบว่าเทคนิคการจัดกลุ่ม จะทำหน้าที่เป็นกลไกในการลดมิติของปัญหา (Problem Complexity Reduction) โดยจะทำการแปลงข้อมูลจำนวนมากให้กลายเป็นโครงสร้างเชิงนามธรรมที่สามารถจัดการได้ง่ายขึ้น การลดความซับซ้อนในลักษณะนี้จึงมีความสำคัญอย่างยิ่งต่ออัลกอริทึมที่ต้องประมวลผลข้อมูลขนาดใหญ่แบบเรียลไทม์ เช่น ระบบค้นคืนข้อมูล ระบบวิเคราะห์เครือข่าย และระบบปัญญาประดิษฐ์ [9] [15] ดังนั้นถ้าหากพิจารณาในบริบทของวิทยาการข้อมูล เทคนิคการจัดกลุ่มจึงมิได้ถูกใช้เพียงเพื่อการวิเคราะห์ข้อมูลเชิงสำรวจ หากแต่ถูกนำมาเป็นองค์ประกอบเชิงโครงสร้างของระบบด้วยเช่นกัน ตัวอย่างเช่น การจัดกลุ่มโหนดในระบบกระจาย การจัดกลุ่มคำหรือเอกสารในระบบค้นคืนข้อมูล และการจัดกลุ่มพฤติกรรมผู้ใช้ในระบบแนะนำอัจฉริยะ เป็นต้น ซึ่งจากที่กล่าวมาทั้งหมดเป็นการทำงานที่ล้วนแต่เป็นการเพิ่มประสิทธิภาพการคำนวณและการจัดสรรทรัพยากรของระบบให้มีความเหมาะสม

เทคนิคการจัดกลุ่มสามารถจำแนกออกเป็นหลายประเภทตามหลักการและวิธีการคำนวณ โดยแต่ละประเภทมีคุณลักษณะเชิงทฤษฎีและข้อจำกัดที่แตกต่างกัน ทั้งนี้พอสรุปได้ว่าเทคนิคการจัดกลุ่มจำแนกออกเป็นดังนี้ การจัดกลุ่มแบบแบ่งส่วน (Partition-based Clustering) การจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering) การจัดกลุ่มแบบอาศัยความหนาแน่น (Density-based Clustering) และการจัดกลุ่มแบบอาศัยแบบจำลอง (Model-based Clustering) [8] เป็นต้น ทั้งนี้เทคนิคการจัดกลุ่มแบบแบ่งส่วน ที่เรียกว่า เทคนิค K-means เป็นวิธีที่ได้รับความนิยมในทางวิทยาการข้อมูลเนื่องจากมีความเรียบง่ายและ

ประสิทธิภาพเชิงคำนวณสูง อย่างไรก็ตาม วิธีดังกล่าวมีข้อจำกัดด้านการกำหนดจำนวนกลุ่มล่วงหน้าและความไวต่อการกำหนดค่าเริ่มต้น ในขณะที่การจัดกลุ่มแบบลำดับขั้นสามารถสะท้อนโครงสร้างเชิงลำดับของข้อมูลได้ดีกว่า แต่มีต้นทุนเชิงคำนวณสูงและไม่เหมาะกับข้อมูลขนาดใหญ่ ในขณะเดียวกันการจัดกลุ่มแบบอาศัยความหนาแน่น เช่น DBSCAN จะมีความสามารถในการตรวจจบบางกลุ่มที่มีรูปร่างซับซ้อนและจัดการกับข้อมูลรบกวนได้ดี แต่มีข้อจำกัดในการกำหนดพารามิเตอร์ที่เหมาะสม ส่วนการจัดกลุ่มแบบอาศัยแบบจำลอง เช่น Gaussian Mixture Models มีความยืดหยุ่นเชิงสถิติสูง แต่ต้องอาศัยสมมติฐานเกี่ยวกับการแจกแจงของข้อมูล ซึ่งอาจไม่สอดคล้องกับข้อมูลจริงในระบบขนาดใหญ่ หากแต่ทั้งนี้ยังสามารถจำแนกเทคนิคการจัดกลุ่มออกเป็นสองลักษณะคือ การจัดกลุ่มแบบ Soft Clustering และการจัดกลุ่มแบบ Hard Clustering ซึ่งการจัดกลุ่มแบบ Soft Clustering เป็นแนวคิดที่ขยายจากการจัดกลุ่มแบบ Hard Clustering โดยอนุญาตให้ข้อมูลหนึ่งรายการสามารถมีค่าสมาชิกในหลายกลุ่มพร้อมกันในระดับความน่าจะเป็นที่แตกต่างกัน ซึ่งแนวคิดดังกล่าวนี้มีความสำคัญอย่างยิ่งต่อการออกแบบอัลกอริทึมที่ต้องจัดการกับความไม่แน่นอนและความซ้อนทับของข้อมูล [6]

ในเชิงทฤษฎี เทคนิคการจัดกลุ่มแบบ Soft Clustering มีความสัมพันธ์อย่างใกล้ชิดกับแบบจำลองเชิงความน่าจะเป็นและการเรียนรู้เชิงสถิติ เช่น Gaussian Mixture Models ซึ่งมองว่าข้อมูลจะถูกสร้างขึ้นจากการผสมของการแจกแจงข้อมูลหลายชุด ซึ่งแนวคิดนี้ได้กลายเป็นรากฐานสำคัญของสถาปัตยกรรม Mixture of Experts ที่ใช้เป็นกลไกการคำนวณเชิงเงื่อนไขในการเลือกส่วนย่อยของแบบจำลองให้เหมาะสมกับข้อมูลนำเข้า [3] เทคนิคการจัดกลุ่มจึงถือได้ว่าเป็นกระบวนการจัดระเบียบการทำงาน ซึ่งจะช่วยให้สามารถควบคุมความซับซ้อนที่เกิดจากการมีองค์ประกอบจำนวนมากและความสัมพันธ์ที่ไม่เป็นเชิงเส้น [10] การแบ่งข้อมูลออกเป็นกลุ่มย่อยจะช่วยลดภาระในการประมวลผลเชิงรวมศูนย์และเพิ่มความสามารถในการปรับตัวของระบบปัญญาประดิษฐ์ โดยเฉพาะอย่างยิ่งในระบบคอมพิวเตอร์ขนาดใหญ่ เช่น ระบบคลาวด์ เป็นต้น จากที่กล่าวมาจะเห็นได้ว่าเทคนิคการจัดกลุ่มมักจะถูกนำมาใช้ในการจัดสรรทรัพยากร การกระจายโหลด และการเพิ่มเสถียรภาพของระบบ ซึ่งหลักการนี้มีความสอดคล้องกับการออกแบบสถาปัตยกรรมแบบกระจาย ซึ่งเป็นพื้นฐานสำคัญของระบบปัญญาประดิษฐ์ยุคใหม่ [11] และหากพิจารณาในบริบทของการออกแบบสถาปัตยกรรมปัญญาประดิษฐ์แล้ว เทคนิคการจัดกลุ่มไม่ได้จำกัดอยู่เพียงในระดับข้อมูลเพื่อทำการพยากรณ์หรือทำนายเท่านั้น หากแต่ขยายไปสู่ระดับการออกแบบแบบจำลองและอัลกอริทึมต่าง ๆ ด้วย ตัวอย่างเช่น การแบ่งโมดูลของแบบจำลอง การแยกหน้าที่ของเครือข่ายประสาทเทียม และการเลือกเส้นทางการประมวลผลที่เหมาะสมกับข้อมูลแต่ละประเภท ดังนั้นจึงสะท้อนให้เห็นว่าการจัดกลุ่มทำหน้าที่เป็นหลักการเชิงสถาปัตยกรรม (Architectural Principle) ที่ช่วยกำหนดโครงสร้างและพฤติกรรมของระบบปัญญาประดิษฐ์ในระดับสูง ซึ่งเป็นพื้นฐานสำคัญของสถาปัตยกรรม Mixture of Experts และระบบปัญญาประดิษฐ์อย่าง DeepSeek ในเวลาต่อมา

### เทคนิคการจัดกลุ่มกับการเรียนรู้การแทนความหมาย (Representation Learning) ในโครงข่ายประสาทเทียมเชิงลึก

การเรียนรู้การแทนความหมาย (Representation Learning) เป็นแนวคิดหลักของการเรียนรู้เชิงลึกที่มุ่งเน้นให้เครื่องจักรสามารถเรียนรู้คุณลักษณะเชิงนามธรรมของข้อมูลจากข้อมูลดิบได้โดยอัตโนมัติ แทนการออกแบบคุณลักษณะด้วยมนุษย์

ซึ่งมีบทบาทสำคัญต่อความสำเร็จของระบบปัญญาประดิษฐ์ในหลากหลายสาขา เช่น การประมวลผลภาษาธรรมชาติ การรู้จำภาพ และการวิเคราะห์สัญญาณ [6] ทั้งนี้ในบริบทของแบบจำลองภาษาขนาดใหญ่ การแทนความหมายเชิงลึกช่วยให้แบบจำลองสามารถจับความสัมพันธ์เชิงบริบท ความหมายเชิงนัย และโครงสร้างเชิงไวยากรณ์ของภาษาได้อย่างมีประสิทธิภาพ การแทนความหมายดังกล่าวมักอยู่ในรูปของเวกเตอร์มิติสูงที่สะท้อนคุณสมบัติหลากหลายของข้อมูล ซึ่งเป็นพื้นฐานสำคัญของกระบวนการเรียนรู้ขั้นถัดไป อย่างไรก็ตาม การแทนความหมายที่มีมิติสูงและความซับซ้อนมากขึ้นย่อมนำไปสู่ภาระในการประมวลผลและการจัดสรรทรัพยากรของเครื่องจักร โดยเฉพาะในเครื่องจักรที่ต้องรองรับข้อมูลและผู้ใช้งานจำนวนมาก ซึ่งปัญหาดังกล่าวเป็นแรงผลักดันให้เกิดการผานแนวคิดการจัดกลุ่มเข้ากับการเรียนรู้การแทนความหมายเพื่อควบคุมความซับซ้อนของเครื่องจักร โดยเมื่อข้อมูลถูกแปลงให้อยู่ในรูปของการแทนความหมายเชิงเวกเตอร์ การจัดกลุ่มก็สามารถถูกนำมาใช้เพื่อค้นหาโครงสร้างเชิงความหมายที่ซ่อนอยู่ การจัดกลุ่มในระดับนี้ช่วยให้ระบบสามารถแบ่งข้อมูลออกเป็นกลุ่มที่มีลักษณะเชิงความหมายใกล้เคียงกัน ซึ่งมีผลต่อการลด search space และเพิ่มประสิทธิภาพของกระบวนการเรียนรู้ [8] จากที่กล่าวมาพบว่ามีงานวิจัยจำนวนมากแสดงให้เห็นว่าเทคนิคการจัดกลุ่มบนการแทนความหมายที่ได้จากโครงข่ายประสาทเทียมเชิงลึกสามารถให้ผลลัพธ์ที่มีคุณภาพสูงกว่าเทคนิคการจัดกลุ่มบนข้อมูลดิบ เนื่องจากการแทนความหมายเชิงลึกสามารถกรองข้อมูลที่ไม่เกี่ยวข้องได้ดีและเน้นคุณลักษณะที่สำคัญต่อการเรียนรู้ของเครื่องจักร [8] ทั้งนี้ในระบบประมวลผลภาษาธรรมชาติ เทคนิคการจัดกลุ่มของเวกเตอร์ที่เป็นคำหรือเวกเตอร์ที่เป็นประโยคสามารถสะท้อนความสัมพันธ์เชิงความหมาย เช่น คำพ้องความหมาย กลุ่มหัวข้อ หรือโครงสร้างเชิงไวยากรณ์ ซึ่งเป็นประโยชน์ต่อการออกแบบระบบค้นคืนข้อมูลและระบบแนะนำอัจฉริยะ

ดังนั้นเทคนิคการจัดกลุ่มในระบบปัญญาประดิษฐ์สมัยใหม่จึงมิได้เป็นกระบวนการคงที่ หากแต่มีลักษณะเชิงพลวัตที่สามารถปรับเปลี่ยนไปตามการเรียนรู้ของเครื่องจักร การจัดกลุ่มเชิงพลวัต (Dynamic Clustering) จะช่วยให้ระบบปัญญาประดิษฐ์สามารถปรับโครงสร้างกลุ่มของข้อมูลให้สอดคล้องกับการเปลี่ยนแปลงข้อมูลและสภาพแวดล้อม ซึ่งมีความสำคัญอย่างยิ่งในระบบปัญญาประดิษฐ์ที่ต้องรองรับข้อมูลแบบสตรีมและการใช้งานจริงในระยะยาว จะเห็นได้ว่าการจัดกลุ่มเชิงพลวัตมีความสัมพันธ์กับการเรียนรู้แบบปรับตัวและการเรียนรู้แบบออนไลน์ ซึ่งช่วยให้ระบบสามารถรักษาสมรรถนะได้แม้เผชิญกับการเปลี่ยนแปลงของรูปแบบข้อมูล [3]

### สถาปัตยกรรม Mixture of Experts (MoE) ในมุมมองเชิงโครงสร้างและเชิงคำนวณ

สถาปัตยกรรม Mixture of Experts (MoE) ถูกนำมาใช้เพื่อตอบสนองต่อข้อจำกัดของรูปแบบแบบหนาแน่น (Dense Models) ซึ่งรูปแบบนี้จะเปิดใช้งานพารามิเตอร์ทั้งหมดในทุกขั้นตอนการประมวลผล แม้ในกรณีที่ข้อมูลนำเข้ามีความซับซ้อนเพียงบางส่วน [3] โดยรูปแบบดังกล่าวนำไปสู่การใช้ทรัพยากรการคำนวณเกินความจำเป็น โดยเฉพาะในแบบจำลองภาษาขนาดใหญ่ที่มีจำนวนพารามิเตอร์ในระดับมหาศาล สถาปัตยกรรมแบบ MoE ที่ถูกนำมาใช้แทนนั้นได้เสนอแนวคิดการแบ่งรูปแบบออกเป็นหน่วยย่อยที่เรียกว่า “ผู้เชี่ยวชาญ” (Experts) ซึ่งแต่ละหน่วยถูกออกแบบให้มีความเชี่ยวชาญเฉพาะด้าน โดยในแต่ละขั้นตอนการประมวลผล ระบบจะเลือกใช้งานผู้เชี่ยวชาญเพียงบางส่วนผ่านกลไกการเลือกเส้นทาง (Routing หรือ Gating)

ซึ่งวิธีนี้จะช่วยให้รูปแบบสามารถเพิ่มจำนวนพารามิเตอร์ได้โดยไม่เพิ่มต้นทุนการคำนวณในสัดส่วนเดียวกัน [12] ดังนั้นสถาปัตยกรรม MoE จึงมีความสัมพันธ์อย่างใกล้ชิดกับการจัดกลุ่มแบบ Soft Clustering เนื่องจากกลไก Gating ทำหน้าที่คล้ายตัวจัดกลุ่มที่กำหนดความน่าจะเป็นในการส่งข้อมูลไปยังผู้เชี่ยวชาญแต่ละราย ซึ่งสะท้อนการแบ่งข้อมูลออกเป็นกลุ่มเชิงฟังก์ชันภายในโมเดล

โดยทั่วไป สถาปัตยกรรม MoE จะประกอบด้วยองค์ประกอบหลัก 3 ส่วน ได้แก่ (1) ชุดผู้เชี่ยวชาญ (Expert Networks) (2) กลไกการเลือกเส้นทางหรือ Gating Network และ (3) กลไกการรวมผลลัพธ์ (Aggregation Mechanism) ซึ่งชุดผู้เชี่ยวชาญมักถูกออกแบบเป็นเครือข่ายประสาทเทียมย่อยที่มีโครงสร้างคล้ายกัน แต่ได้รับการฝึกฝนให้ตอบสนองต่อรูปแบบข้อมูลที่แตกต่างกัน ในขณะที่ กลไกการเลือกเส้นทาง ทำหน้าที่ประเมินข้อมูลนำเข้าและกำหนดน้ำหนักหรือความน่าจะเป็นในการเลือกผู้เชี่ยวชาญแต่ละราย และกลไกการรวมผลลัพธ์จากผู้เชี่ยวชาญที่ถูกเลือกจะถูกนำมาประกอบเป็นผลลัพธ์สุดท้ายของรูปแบบ อาจกล่าวได้ว่าสถาปัตยกรรม MoE จะช่วยให้การประมวลผลมีลักษณะเชิงเงื่อนไข (Conditional Computation) ซึ่งเป็นหัวใจสำคัญของปัญญาประดิษฐ์ ลดภาระการคำนวณในระบบขนาดใหญ่ได้เป็นอย่างดี [3] ซึ่งเมื่อศึกษางานวิจัยด้าน สถาปัตยกรรม MoE ทำให้สามารถจำแนกสถาปัตยกรรมออกเป็นหลายประเภทตามลักษณะของกลไกการเลือกเส้นทางและระดับในการคำนวณ โดยสามารถสรุปเป็นกลุ่มหลัก ได้แก่ MoE แบบดั้งเดิม, Sparse MoE และ Conditional MoE เพื่อให้เห็นภาพความแตกต่างของสถาปัตยกรรม MoE จึงสรุปได้ดังนี้ [12]

**ตารางที่ 1** การจำแนกประเภทของสถาปัตยกรรม Mixture of Experts

| ประเภท MoE      | ลักษณะเด่น                     | กลไกการเลือกเส้นทาง | ภาระการคำนวณ |
|-----------------|--------------------------------|---------------------|--------------|
| MoE แบบดั้งเดิม | ใช้ผู้เชี่ยวชาญหลายรายพร้อมกัน | Soft Gating         | สูง          |
| Sparse MoE      | เลือกผู้เชี่ยวชาญเพียงบางส่วน  | Top-k Gating        | ปานกลาง      |
| Conditional MoE | คำนวณตามเงื่อนไขของข้อมูล      | Learned Routing     | ต่ำ          |

กลไก Gating ทำหน้าที่เป็นศูนย์กลางของสถาปัตยกรรม MoE โดยกำหนดว่าข้อมูลนำเข้าควรถูกส่งไปประมวลผลโดยผู้เชี่ยวชาญรายใด กลไกนี้สามารถตีความได้ว่าเป็นกระบวนการจัดกลุ่มเชิงฟังก์ชัน (Functional Clustering) ที่จัดกลุ่มข้อมูลตามลักษณะการประมวลผลที่เหมาะสม การออกแบบกลไก Gating ที่มีประสิทธิภาพจึงต้องคำนึงถึงทั้งความแม่นยำในการเลือกผู้เชี่ยวชาญและความสมดุลของโหลดระหว่างผู้เชี่ยวชาญต่าง ๆ จากการศึกษพบว่าปัญหาความไม่สมดุลของโหลดเป็นหนึ่งในความท้าทายสำคัญของสถาปัตยกรรม MoE ซึ่งต้องอาศัยเทคนิคการปรับสมดุลเพิ่มเติม [3] เพื่อให้เห็นภาพชัดเจนยิ่งขึ้นจึงเปรียบเทียบสถาปัตยกรรมระหว่าง Dense Transformer และ MoE Transformer ดังตารางด้านล่างนี้

ตารางที่ 2 การเปรียบเทียบ Dense Transformer และ MoE Transformer

| รายการ                  | Dense Transformer  | MoE Transformer |
|-------------------------|--------------------|-----------------|
| การใช้พารามิเตอร์       | ใช้ทั้งหมดทุกครั้ง | ใช้บางส่วน      |
| รูปแบบการคำนวณ          | ซับซ้อน            | ไม่ซับซ้อน      |
| ความสามารถในการขยายขนาด | ขยายขนาดมีขีดจำกัด | ขยายขนาดได้มาก  |
| การจัดสรรทรัพยากร       | คงที่              | ปรับตามข้อมูล   |
| ความซับซ้อนเชิงระบบ     | ต่ำกว่า            | สูงกว่า         |

จากตารางที่ 2 จะเห็นได้ว่าสถาปัตยกรรม MoE ช่วยเพิ่มความสามารถในการขยายขนาดของรูปแบบโดยไม่เพิ่มภาระการคำนวณในระดับเดียวกัน แต่อย่างไรก็ตามความซับซ้อนเชิงระบบที่ใช้ในการคำนวณที่เพิ่มขึ้นจำเป็นต้องได้รับการจัดการที่เหมาะสมในระดับสถาปัตยกรรม ทั้งนี้เมื่อพิจารณาในภาพรวม จะเห็นได้ว่าสถาปัตยกรรม MoE เมื่อนำมาใช้งานร่วมกับเทคนิคการจัดกลุ่มจะเกิดการทำงานได้อย่างมีประสิทธิภาพ ซึ่งหลักการนี้ช่วยให้ระบบสามารถควบคุมความซับซ้อนเชิงคำนวณเพิ่มประสิทธิภาพการใช้ทรัพยากร และสนับสนุนการออกแบบระบบปัญญาประดิษฐ์ที่สามารถขยายขนาดได้อย่างยั่งยืน ซึ่งเป็นพื้นฐานสำคัญของระบบปัญญาประดิษฐ์อย่าง DeepSeek นั่นเอง

### แนวคิดการคำนวณแบบ Sparse และการคำนวณเชิงเงื่อนไข

ข้อจำกัดประการหนึ่งของการคำนวณแบบ Dense Computation ในระบบคอมพิวเตอร์และระบบสารสนเทศสมัยใหม่คือต้องรองรับปริมาณข้อมูลและภาระงานที่เพิ่มขึ้นอย่างต่อเนื่อง การคำนวณแบบ Dense Computation จึงต้องเปิดใช้งานทรัพยากรทั้งหมดของระบบคอมพิวเตอร์เพื่อการประมวลผล โดยจะแสดงข้อจำกัดให้เห็นอย่างชัดเจน ทั้งในด้านเวลาหน่วยความจำ และพลังงาน ส่งผลให้การเพิ่มขนาดของรูปแบบหรือระบบเพิ่มขึ้นแบบทวีคูณ ซึ่งไม่สอดคล้องกับข้อจำกัดของโครงสร้างพื้นฐานทางคอมพิวเตอร์ในทางปฏิบัติ [11] ดังนั้นในบริบทของปัญญาประดิษฐ์ โดยเฉพาะในประเด็นของแบบจำลองภาษาขนาดใหญ่ การใช้การคำนวณแบบ Dense Computation ทำให้ทุกพารามิเตอร์ต้องถูกประมวลผลในทุกคำหรือทุกโทเคน แม้ข้อมูลนำเข้าจะต้องการเพียงความสามารถเฉพาะบางส่วนของรูปแบบเท่านั้น จากปัญหาดังกล่าวจึงเป็นแรงผลักดันให้เกิดแนวคิดการคำนวณแบบ Sparse และการคำนวณเชิงเงื่อนไขขึ้น โดยการคำนวณแบบ Sparse Computation ในเชิงสถาปัตยกรรมคือการคำนวณที่จะกระจายองค์ประกอบออกเป็นส่วนย่อย ๆ เพื่อลดจำนวนองค์ประกอบเหล่านั้นลง และเปิดใช้งานในเฉพาะที่ต้องการในแต่ละรอบการประมวลผล โดยเลือกใช้เฉพาะส่วนของระบบที่มีความจำเป็นต่อข้อมูลนำเข้าในขณะนั้น ทั้งนี้แนวคิดนี้มีบทบาทสำคัญในหลายระดับของระบบคอมพิวเตอร์ ตั้งแต่ระดับฮาร์ดแวร์ เช่น การจัดการหน่วยความจำและแคช ไปจนถึงระดับซอฟต์แวร์และอัลกอริทึม และเมื่อมองลงไปในระดับสถาปัตยกรรมของระบบปัญญาประดิษฐ์ การคำนวณแบบ Sparse ถูกนำมาใช้เพื่อควบคุมภาระการคำนวณของรูปแบบขนาดใหญ่ โดยบูรณาการร่วมกับสถาปัตยกรรม Mixture of Experts และจุดนี้เองจะเห็นได้ว่าเป็นการประยุกต์ใช้ในระดับแบบจำลอง ที่จะช่วยให้ระบบปัญญาประดิษฐ์สามารถเพิ่มจำนวนพารามิเตอร์ได้โดยไม่เพิ่มต้นทุนการคำนวณในอัตราส่วนเดียวกัน [12]

การคำนวณเชิงเงื่อนไข (Conditional Computation) เป็นแนวคิดที่ต่อยอดจากการคำนวณแบบ Sparse Computation โดยเป็นการคำนวณหรือประมวลผลที่เน้นการตัดสินใจเชิงเงื่อนไข ว่าควรเปิดใช้งานส่วนใดของระบบในแต่ละสถานการณ์ ซึ่งหลักการในประเด็นนี้มีความสอดคล้องกับการออกแบบระบบคอมพิวเตอร์เชิงปรับตัว ที่สามารถตอบสนองต่อลักษณะของภาระงานที่เปลี่ยนแปลงไปได้อย่างมีประสิทธิภาพ ทั้งนี้ในสถาปัตยกรรม MoE ที่มีองค์ประกอบในประเด็นของกลไก Gating หรือการเลือกเส้นทาง ที่จะทำหน้าที่เป็นกลไกการตัดสินใจเชิงเงื่อนไขที่เลือกผู้เชี่ยวชาญที่เหมาะสมกับข้อมูลนำเข้าในแต่ละขั้นตอน การเลือกเส้นทางดังกล่าวช่วยลดภาระการคำนวณโดยรวมและเพิ่มประสิทธิภาพของระบบ โดยเฉพาะในระบบที่ต้องรองรับการประมวลผลพร้อมกันจำนวนมากได้ [3] นอกจากหลักการที่กล่าวมาแล้ว 2 ประการ ยังอาจต้องให้แนวคิดในเรื่องของ เทคนิคการจัดกลุ่มเข้ามาผสมผสานกับการคำนวณหรือการใช้ทรัพยากรอย่างมีประสิทธิภาพอีกด้วย ซึ่งเทคนิคดังกล่าวคือ เทคนิคการจัดกลุ่มเชิงเป้าหมาย (Targeted Clustering) ซึ่งเป็นแนวคิดที่ผสมผสานการจัดกลุ่มเข้ากับการคำนวณเชิงเงื่อนไขเข้าด้วยกัน โดยมุ่งเน้นการจัดกลุ่มข้อมูลหรือภาระงานตามลักษณะการใช้ทรัพยากรและความต้องการในการประมวลผล ซึ่งแนวคิดนี้จะช่วยให้ระบบปัญญาประดิษฐ์สามารถจัดสรรทรัพยากรได้อย่างเหมาะสม ลดความซ้ำซ้อน และเพิ่มอัตราการใช้ทรัพยากรให้เกิดประโยชน์สูงสุด ทั้งนี้ในบริบทของสถาปัตยกรรม MoE การจัดกลุ่มเชิงเป้าหมายช่วยให้ผู้เชี่ยวชาญแต่ละรายถูกออกแบบและปรับแต่งให้เหมาะสมกับกลุ่มข้อมูลเฉพาะ ส่งผลให้ระบบสามารถตอบสนองต่อข้อมูลที่มีความหลากหลายได้อย่างมีประสิทธิภาพมากขึ้น แนวคิดนี้มีบทบาทสำคัญต่อการออกแบบระบบปัญญาประดิษฐ์ที่ต้องการทั้งสมรรถนะและความยั่งยืนเชิงระบบ

จากแนวคิด วิธีการ หรือหลักการที่กล่าวมาทั้งหมดในข้างต้นล้วนแต่เป็นการนำไปใช้ในการออกแบบโครงสร้างและสถาปัตยกรรมของระบบปัญญาประดิษฐ์ DeepSeek [13] [16] [17] กล่าวคือ เมื่อเชื่อมโยงแนวคิดการคำนวณแบบ Sparse และการคำนวณเชิงเงื่อนไขเข้าไปในระบบปัญญาประดิษฐ์ DeepSeek จะทำให้การทำงานที่รองรับความซับซ้อนของข้อมูลขนาดใหญ่รองรับภาษาได้หลากหลาย ลดต้นทุนการคำนวณและเพิ่มความสามารถในการตอบสนองต่อการใช้งานจริงนอกจากนี้ ยังแสดงให้เห็นว่าเทคนิคการจัดกลุ่ม และสถาปัตยกรรม MoE มีใช้เพียงเทคนิคเชิงอัลกอริทึม หากแต่เป็นหลักการออกแบบระบบคอมพิวเตอร์ที่มีนัยสำคัญต่อการพัฒนาระบบปัญญาประดิษฐ์ DeepSeek ในเวอร์ชันรุ่นใหม่ ๆ ในอนาคต

### บทบาทของเทคนิคการจัดกลุ่ม และ สถาปัตยกรรม Mixture of Experts ในระบบปัญญาประดิษฐ์ DeepSeek

DeepSeek เป็นระบบปัญญาประดิษฐ์ที่ได้รับความนิยมอย่างมากในกลุ่มผู้ใช้งาน และยิ่งได้รับความนิยมมากขึ้นเมื่อ DeepSeek ได้นำเสนอแนวทางการออกแบบแบบจำลองภาษาขนาดใหญ่ที่เน้นประสิทธิภาพเชิงระบบ ควบคู่กับใช้วิธีการเรียนรู้ของเครื่องจักรแบบเทคนิคการจัดกลุ่ม แม้ว่ารายละเอียดเชิงลึกบางส่วนจากระบบจะไม่ถูกเปิดเผยอย่างสมบูรณ์ เนื่องจากเป็นความลับทางการค้า และเพื่อป้องกันการแย่งความเป็นผู้นำด้านเทคโนโลยีจากคู่แข่ง แต่จากการสืบค้น ค้นคว้า และการวิเคราะห์เอกสาร พบว่า DeepSeek ใช้แนวคิดการออกแบบที่สอดคล้องกับสถาปัตยกรรมแบบ Sparse และการคำนวณเชิงเงื่อนไข (ซึ่งมีรากฐานมาจากเทคนิคการจัดกลุ่ม และ สถาปัตยกรรม Mixture of Experts (MoE) ) และหากพิจารณาจากมุมมองเชิงวิศวกรรมคอมพิวเตอร์ จะพบอีกว่า DeepSeek ไม่ได้มุ่งเน้นการเพิ่มจำนวนพารามิเตอร์เพียงอย่างเดียว หากแต่ให้ความสำคัญกับ

การควบคุมการใช้ทรัพยากร การกระจายโหนดการคำนวณ และความสามารถในการปรับขนาดของระบบ ซึ่งเป็นประเด็นสำคัญในการพัฒนาระบบปัญญาประดิษฐ์

หากแต่อาจมีคำถามว่า แล้วทำไมระบบปัญญาประดิษฐ์ DeepSeek จึงใช้แนวคิดเทคนิคการจัดกลุ่มและ สถาปัตยกรรมแบบ Mixture of Experts มาเป็นโครงสร้างหลัก และเมื่อพิจารณาแล้วจะพบว่า เทคนิคการจัดกลุ่มในระบบ DeepSeek สามารถตีความได้ในหลายระดับ ตั้งแต่ระดับการแทนความหมายของข้อมูล ไปจนถึงระดับการจัดกลุ่มฟังก์ชันการประมวลผลภายในแบบจำลอง เทคนิคการจัดกลุ่มยังช่วยให้ระบบสามารถแยกข้อมูลที่มีลักษณะเชิงความหมายแตกต่างกันออกจากกัน [14] และเลือกใช้ทรัพยากรการคำนวณที่เหมาะสมกับข้อมูลแต่ละประเภทในระดับสถาปัตยกรรม และที่สำคัญที่สุดคือเทคนิคการจัดกลุ่มมีบทบาทในการกำหนดโครงสร้างของผู้เชี่ยวชาญในสถาปัตยกรรม MoE โดยผู้เชี่ยวชาญแต่ละรายสามารถถูกมองว่าเป็นตัวแทนของกลุ่มฟังก์ชันการประมวลผลที่แตกต่างกัน การออกแบบในลักษณะนี้ช่วยลดความซ้ำซ้อนของการคำนวณและเพิ่มประสิทธิภาพของระบบโดยรวม จากที่กล่าวมา DeepSeek จึงใช้แนวคิดสถาปัตยกรรมแบบ Mixture of Experts เพื่อสนับสนุนการคำนวณเชิงเงื่อนไข โดยกลไกการเลือกเส้นทางทำหน้าที่กำหนดว่าข้อมูลนำเข้าควรถูกส่งไปประมวลผลโดยผู้เชี่ยวชาญรายใดในแต่ละขั้นตอน แนวคิดดังกล่าวช่วยให้ระบบสามารถใช้พารามิเตอร์จำนวนมากโดยไม่เพิ่มภาระการคำนวณในระดับเดียวกัน และยังช่วยให้การขยายขนาดของโมเดลเป็นไปอย่างมีประสิทธิภาพ โดยสามารถเพิ่มจำนวนผู้เชี่ยวชาญเพื่อรองรับความหลากหลายของข้อมูลและภาระงานได้โดยไม่ส่งผลกระทบต่ออย่างมีนัยสำคัญต่อเวลาในการตอบสนองของระบบ

จากกรณีของ ระบบปัญญาประดิษฐ์ DeepSeek ถือได้ว่าเป้าหมายความสำเร็จของระบบปัญญาประดิษฐ์ ที่ไม่ได้เกิดจากการเลือกใช้เทคโนโลยีใดเทคโนโลยีหนึ่งโดยลำพัง หากแต่เกิดจากการผสมผสานแนวคิดเชิงทฤษฎีและเชิงสถาปัตยกรรมอย่างเป็นระบบ อย่างเช่น เทคนิคการจัดกลุ่มและสถาปัตยกรรมแบบ Mixture of Experts ที่ทำหน้าที่เป็นกลไกหลักในการควบคุมความซับซ้อน เพิ่มประสิทธิภาพการคำนวณ และสนับสนุนการขยายขนาดของระบบ ระบบปัญญาประดิษฐ์ DeepSeek สะท้อนให้เห็นแนวโน้มการออกแบบระบบปัญญาประดิษฐ์ที่มุ่งเน้นความยั่งยืนเชิงระบบ มากกว่าการแข่งขันด้านขนาดของแบบจำลองเพียงอย่างเดียว ดังนั้นบทบาทของเทคนิคการจัดกลุ่ม และสถาปัตยกรรมแบบ Mixture of Experts จึงอยู่ในฐานะกลไกสำคัญที่อยู่เบื้องหลังความสำเร็จของระบบ DeepSeek ผ่านมุมมองเชิงทฤษฎีและเชิงสถาปัตยกรรม ผลการวิเคราะห์เอกสารชี้ให้เห็นว่าแนวคิดดังกล่าวมีบทบาทสำคัญต่อการควบคุมความซับซ้อน การเพิ่มประสิทธิภาพการคำนวณ และการออกแบบระบบปัญญาประดิษฐ์ที่สามารถขยายขนาดได้อย่างยั่งยืน และข้อค้นพบเชิงแนวคิดจากบทความนี้อาจจะใช้เป็นกรอบอ้างอิงสำหรับการออกแบบและพัฒนาระบบปัญญาประดิษฐ์และระบบคอมพิวเตอร์สมัยใหม่ในอนาคต โดยเฉพาะในบริบทของการจัดการข้อมูลและทรัพยากรในระดับขนาดใหญ่

## เอกสารอ้างอิง

- [1] A. Vaswani et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.
- [2] T. B. Brown et al., “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [3] N. Shazeer et al., “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.
- [4] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.
- [5] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” *arXiv preprint arXiv:2108.07258*, 2021. [Online]. Available: <https://arxiv.org/abs/2108.07258>. [Accessed: Feb. 2, 2026].
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [8] A. K. Jain, “Data clustering: 50 years beyond k-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.
- [9] T. Zhang, H. Guo, W. Lu, T. Dai, S.-T. Xia, and J. Wang, “SPARSEEVAL: Efficient Evaluation of Large Language Models by Sparse Optimization,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2026. [Online]. Available: <https://openreview.net/forum?id=CZAzAedGSV>. [Accessed: Feb. 10, 2026].
- [10] M. Mitchell, *Complexity: A Guided Tour*. Oxford, U.K.: Oxford University Press, 2009.
- [11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2020.
- [12] W. Fedus, B. Zoph, and N. Shazeer, “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity,” *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
- [13] S. Xu, “The Complete Guide to DeepSeek Models: V3, R1, V4 and Beyond,” *BentoML Blog*, Apr. 24, 2026. [Online]. Available: <https://www.bentoml.com/blog/the-complete-guide-to-deepseek-models-from-v3-to-r1-and-beyond>. [Accessed: Feb. 10, 2026].

- 
- [14] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," in *Proc. 6th Symp. Operating Systems Design and Implementation (OSDI)*, San Francisco, CA, USA, Dec. 2004, pp. 137–150.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [16] DeepSeek-AI, "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model," *arXiv preprint arXiv:2405.04434*, 2024. [Online]. Available: <https://arxiv.org/abs/2405.04434>. [Accessed: Feb. 2, 2026].
- [17] D. Dai *et al.*, "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Bangkok, Thailand, Aug. 2024, pp. 1280–1297, doi: 10.18653/v1/2024.acl-long.70.