George Box名言對LLM發展的啟發性意義

George Box的經典名言「All models are wrong, but some are useful」(所有模型都是錯誤的，但有些是有用的)對現今大型語言模型(LLM)的發展具有深刻的啟發意義，特別是在處理模型幻覺和對現實世界理解的局限性方面。

Box哲學的核心洞察

Box的這句話反映了一個重要的認識論觀點：任何模型都是對複雜現實的簡化抽象，因此必然存在不完美之處。這種「認識論謙遜」(epistemic humility)承認模型的內在局限性，同時強調其實用價值取決於特定應用情境下的有用性，而非絕對的準確性。[1][2][3]

對LLM幻覺現象的解釋框架

幻覺的本質性特徵

從Box的哲學視角來看，LLM的幻覺現象並非純粹的技術缺陷，而是模型本質的必然結果。研究顯示，幻覺是生成式AI的「內在特徵」，因為模型只是對訓練數據中詞語關係的學習抽象，而非真實知識的完整儲存。這呼應了Box對模型作為「近似」而非「真理」的理解。[4][2][5]

不同類型的幻覺

最新研究將LLM幻覺分為兩類：HK-(模型缺乏所需知識)和HK+(模型擁有知識但仍產生錯誤答案)。這種區分體現了Box哲學中的重要洞察：模型的「錯誤」有不同層次和原因，需要針對性的解決方案。[6][7]

對現實世界理解的認識論意義

模型與現實的根本鴻溝

Box的觀點揭示了一個深層問題：LLM作為語言模式匹配系統，本質上無法真正「理解」現實世界，只能基於統計規律生成看似合理的回應。這種局限性源於「地圖不是領土」的基本原則——任何表徵都無法完全等同於它所描述的現實。[8][9][2]

不確定性與置信度的校準問題

現代LLM面臨嚴重的置信度校準問題，經常對錯誤答案表現出高度確信。這反映了Box哲學中的核心警示：模型的自信程度並不等同於其準確性。研究表明，即使模型擁有正確知識，仍可能以高確信度產生幻覺。[10][11]

對AI發展的指導意義

認識論謙遜的必要性

Box的哲學提倡在AI系統設計中融入「認識論謙遜」的原則。這意味著AI系統應該：

1. 明確承認不確定性：系統應該能夠識別並表達其知識邊界[12][13]
2. 避免過度自信：防止在不確定情況下表現出不當的確信[14][15]
3. 支持人類判斷：將AI視為增強而非替代人類認知的工具[13][16]

設計哲學的轉變

從Box的視角出發，AI開發應該從追求「完美模型」轉向建構「有用的近似」。這包括：

1. 透明度優先：重視模型可解釋性，讓使用者理解其局限性[17][18]
2. 情境適應性：根據不同應用場景調整模型的確信度閾值[15]
3. 持續校準：建立機制來監測和調整模型的置信度校準[19][14]

實踐層面的啟發

減輕幻覺的策略

基於Box的哲學，減輕LLM幻覺的策略應該包括：

1. 多樣化驗證：使用多種方法交叉驗證模型輸出
2. 外部知識整合：結合可靠的外部知識庫進行事實檢驗[9]
3. 不確定性量化：開發更精確的不確定性估計方法[20][15]

使用者教育的重要性

Box的名言強調了使用者理解模型局限性的重要性。這要求：

1. 風險意識培養：幫助使用者認識到AI輸出的潛在不準確性

2. 批判性思維：鼓勵對AI生成內容進行獨立驗證
3. 適當期望設定：建立對AI能力的現實期望

哲學層面的深層意涵

知識的本質問題

Box的觀點觸及了一個根本性的認識論問題：什麼構成了有效的知識表徵？在LLM的語境下，這意味著我們需要重新思考「理解」和「知識」的定義。[21][22]

人機協作的新範式

最終，Box的哲學指向了一種新的人機協作範式，其中AI系統被視為有用但有限的工具，需要與人類判斷相結合才能產生可靠的結果。[16][13]

結論

George Box的名言為理解和改進LLM提供了一個重要的認識論框架。它提醒我們，幻覺並非單純的技術問題，而是模型本質的反映。通過擁抱認識論謙遜，我們可以更好地設計、使用和信任AI系統，同時保持對其局限性的清醒認識。這種平衡的觀點對於AI的安全、可靠發展至關重要，特別是在高風險應用領域中。

⁂

---

1. https://en.wikipedia.org/wiki/All_models_are_wrong
2. https://egtheory.wordpress.com/2013/11/06/wrong-models/
3. https://www.kdnuggets.com/2019/06/all-models-are-wrong.html
4. https://nonodename.com/post/allmodels/
5. https://blogs.sas.com/content/iml/2025/04/02/all-models-are-wrong.html
6. https://arxiv.org/html/2410.22071v2
7. https://arxiv.org/html/2410.22071v1
8. https://arxiv.org/html/2412.17883v1

9. https://climatetrace.org/news/how-climate-trace-guards-against-the-ai-hallucination
10. https://arxiv.org/html/2502.12964v1
11. https://www.nature.com/articles/s42256-024-00976-7
12. https://arxiv.org/html/2509.09658v1
13. https://pmc.ncbi.nlm.nih.gov/articles/PMC10423547/
14. https://arxiv.org/html/2509.01564v1
15. https://arxiv.org/html/2503.15850v1
16. https://www.tandfonline.com/doi/full/10.1080/02691728.2025.2466164
17. https://hyperight.com/ai-black-box-what-were-still-getting-wrong-about-trusting-machine-learning-models/
18. https://www.ibm.com/think/topics/interpretability
19. https://arxiv.org/html/2412.14660v2
20. https://arxiv.org/pdf/2506.17331.pdf
21. https://arxiv.org/html/2506.18852v1
22. https://arxiv.org/html/2408.11441v1
23. https://www.reddit.com/r/MachineLearning/comments/81wrgi/d_what_is_the_difference_between_confidence_and/
24. https://aclanthology.org/2024.emnlp-main.116/
25. https://aclanthology.org/2024.naacl-long.366.pdf
26. https://www.nature.com/articles/s41586-024-07421-0
27. https://dl.acm.org/doi/10.1145/3711896.3736569
28. https://www.sei.cmu.edu/blog/beyond-capable-accuracy-calibration-and-robustness-in-large-language-models/
29. https://softwaremind.com/blog/llm-hallucinations-definition-examples-and-potential-remedies/

30. https://dl.acm.org/doi/10.1145/3703155
31. https://arxiv.org/html/2311.14648v3
32. https://arxiv.org/html/2505.19240v1
33. https://arxiv.org/html/2310.01469v3
34. https://arxiv.org/pdf/2304.00612.pdf
35. https://arxiv.org/html/2502.16143v1
36. https://arxiv.org/html/2406.13138v1
37. https://arxiv.org/pdf/2002.09595.pdf
38. https://arxiv.org/html/2412.04503v1
39. https://arxiv.org/html/2302.03460v3
40. https://arxiv.org/html/2505.19145v1
41. https://arxiv.org/html/2405.06800v1
42. https://arxiv.org/pdf/2310.19819.pdf
43. https://arxiv.org/html/2401.01313v2
44. https://arxiv.org/html/2405.06800v3
45. https://arxiv.org/pdf/2407.15671.pdf
46. https://arxiv.org/html/2311.05232v2
47. https://arxiv.org/pdf/2307.06435.pdf
48. https://snorkel.ai/large-language-models/
49. https://bigohtech.com/hallucinations-in-ai-models-a-quick-guide
50. https://transformer-circuits.pub/2025/attribution-graphs/biology.html
51. https://arxiv.org/html/2407.18782v1
52. https://www.reddit.com/r/learnmachinelearning/comments/18pl1wx/is_it_true_that_current_llms_are_actually_black/

53. https://www.sciencedirect.com/science/article/pii/S2666386425001158

54. https://www.electronicsforu.com/technology-trends/do-not-be-fooled-by-ai-hallucinations

55. https://sloanreview.mit.edu/article/the-working-limitations-of-large-language-models/

56. https://www.cmu.edu/dietrich/philosophy/docs/london/hastings.pdf

57. https://arxiv.org/pdf/2203.11147.pdf

58. https://pmc.ncbi.nlm.nih.gov/articles/PMC11638409/

59. https://www.linkedin.com/posts/brandon-perelman-184a4343_help-needed-missing-scientist-who-are-activity-7320101619231072260-G8Ix

60. https://read.dukeupress.edu/critical-ai/article/doi/10.1215/2834703X-11556011/400182/Eight-Things-to-Know-about-Large-Language-Models

61. https://arxiv.org/html/2412.20485v1

62. https://arxiv.org/html/2506.18183v1

63. https://arxiv.org/pdf/2504.19255.pdf

64. https://arxiv.org/pdf/2311.03533.pdf

65. https://arxiv.org/html/2504.20084v1

66. https://arxiv.org/html/2507.23020v1

67. https://arxiv.org/pdf/2401.14655.pdf

68. https://arxiv.org/pdf/2507.21067.pdf

69. https://arxiv.org/pdf/1203.0952.pdf

70. https://arxiv.org/pdf/2411.10482.pdf

71. https://arxiv.org/pdf/2208.11230.pdf

72. https://arxiv.org/html/2412.15030v1

73. https://arxiv.org/html/2405.05386v1

74. https://arxiv.org/pdf/2007.05748.pdf

75. https://arxiv.org/html/2507.21067v1

76. https://arxiv.org/html/2410.16806v1

77. https://arxiv.org/pdf/1902.05539.pdf

78. https://statmodeling.stat.columbia.edu/2012/03/04/all-models-are-right-most-are-useless/

79. https://iep.utm.edu/models/

80. https://cran.r-project.org/web/packages/statquotes/vignettes/statquotes.html

81. https://www.sciencedirect.com/science/article/pii/S0001691823001555

82. https://www.reddit.com/r/PhilosophyofScience/comments/188s9b1/all_models_are_wrongbut_are_they_though/

83. https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/

84. https://ethicsblog.crb.uu.se/2024/06/04/artificial-consciousness-and-the-need-for-epistemic-humility/

85. https://fs.blog/all-models-are-wrong/

86. https://www.funblocks.net/thinking-matters/classic-mental-models/epistemic-humility

87. https://thestrategybridge.org/the-bridge/2015/9/20/right-wrong-relevant-how-to-be-right-in-the-ways-that-matter

88. https://jme.bmj.com/cgi/content/short/jme-2024-110591v1?rss=1

89. https://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf

90. https://academic.oup.com/edited-volume/59762/chapter/508611708

91. https://statmodeling.stat.columbia.edu/2008/06/12/all_models_are/

92. https://arxiv.org/html/2508.03860v1

93. https://arxiv.org/html/2407.18950v4

94. https://arxiv.org/pdf/2504.18601.pdf

95. https://arxiv.org/pdf/2103.00752.pdf

96. https://arxiv.org/html/2503.02623v1

97. https://arxiv.org/html/2506.11021v1

98. https://openreview.net/pdf?id=MEUIaOlUxY

99. https://arxiv.org/html/2408.05093v1

100. https://www.arxiv.org/pdf/2406.17836.pdf

101. https://arxiv.org/html/2504.02902v1

102. https://arxiv.org/html/2509.04664v1

103. https://openreview.net/pdf?id=r1bWQ4GubB

104. https://www.reddit.com/r/OpenAI/comments/1na1zyf/openai_just_found_cause_of_hallucinations_of/

105. https://arxiv.org/html/2312.16229v1

106. https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf

107. https://revistas.unir.net/index.php/ijimai/article/download/648/763