



Review

for the monograph

Practical Benchmarking and Hands-On Evaluation of AI Popular Systems
Authored by: Prof Dr Alexander Iliev

From:

Prof. Dr. Goran Rafajlovski

*Professor für Energietechnik
School of Technology and Architecture
SRH University of Applied Sciences Heidelberg
Sonnenallee 221 A-F, D-12059 Berlin
Phone: +49 (0) 178 519 5124
Email: goran.rafajlovski@srh.de
Web: www.srh-berlin.de*

SRH University of Applied Sciences, Berlin

Date: 25. June.2026

The rapid evolution of large language models and AI assistants has created a growing need for objective, reproducible, and experimentally grounded evaluation methodologies. While many publications describe the capabilities of modern AI systems from a conceptual perspective, comparatively few provide systematic experimental evidence obtained under consistent benchmarking conditions. The manuscript *Practical Benchmarking and Hands-On Evaluation of AI Popular Systems* addresses this gap by presenting a comprehensive hands-on comparative evaluation of leading AI systems through a unified experimental framework centered on practical implementation tasks. The book complements the theoretical companion volume by demonstrating how different AI models perform when solving identical engineering problems under comparable conditions.

A major strength of the manuscript lies in its emphasis on reproducible experimentation. Rather than relying on subjective impressions of model quality, the author evaluates multiple contemporary AI systems—including ChatGPT 3.5, ChatGPT 4, ChatGPT 5.2, Claude, Gemini, Bing AI, Zapier Chatbot, Meta Llama, Gemma, Mistral, and Hugging Chat—using the same practical deep learning task based on the MNIST benchmark dataset. This consistent evaluation strategy enables meaningful comparisons between systems while minimizing methodological bias introduced by differing experimental settings.

The experimental design is particularly well structured. Each AI system is tasked with generating complete TensorFlow/Keras implementations of convolutional neural



networks, beginning with environment setup and data preprocessing, continuing through model architecture definition, compilation, training, evaluation, visualization, and interpretation of results. The methodology therefore evaluates not only the correctness of generated code, but also the systems' ability to guide complete machine learning workflows from problem formulation to experimental validation.

Another important contribution is the extensive documentation of the interaction process between the researcher and each AI system. Rather than presenting only final performance metrics, the manuscript records prompts, generated code, execution results, graphical outputs, troubleshooting procedures, hyperparameter optimization strategies, and iterative refinements. This provides valuable insight into the practical usability of modern AI assistants for software engineering and machine learning development while substantially improving reproducibility.

The manuscript goes beyond simple code generation by incorporating objective evaluation techniques commonly used in machine learning research. Accuracy curves, loss curves, confusion matrices, ROC analysis, precision-recall curves, feature-map visualization, hyperparameter optimization, learning-rate experiments, activation-function comparisons, K-fold cross-validation, and multiple architectural variations are systematically explored. This transforms the work from a collection of demonstrations into a structured benchmarking study supported by quantitative evidence.

From a scientific perspective, the comparative methodology represents one of the strongest aspects of the book. All AI systems are evaluated using essentially identical objectives, enabling readers to distinguish differences in reasoning quality, code correctness, implementation completeness, optimization capability, debugging behavior, and final model performance. Such standardized benchmarking is considerably more informative than isolated case studies because it enables direct comparison across multiple state-of-the-art systems.

The manuscript also demonstrates excellent practical relevance. Software engineers, AI researchers, educators, and graduate students increasingly rely on generative AI during software development. Understanding not only whether these systems can generate code, but also how reliably they produce executable, optimized, and experimentally valid implementations has become an important research problem. By documenting both successful and problematic outputs, the author presents a balanced and technically credible evaluation that reflects realistic usage scenarios.

The presentation is clear, logically organized, and highly reproducible. The progression from introductory implementations toward increasingly sophisticated optimization experiments allows readers with different backgrounds to follow the development process while also providing sufficient technical depth for advanced practitioners. Numerous code listings, experimental outputs, performance plots, and implementation examples considerably enhance the educational and scientific value of the manuscript.



From the standpoint of originality, the work contributes a systematic experimental comparison of contemporary AI systems under a common practical framework. Although individual AI models have been evaluated separately in numerous publications, comprehensive side-by-side benchmarking using identical deep learning implementation tasks remains relatively uncommon. The manuscript therefore provides an original contribution by combining prompt engineering, software implementation, experimental validation, and comparative AI evaluation within a unified methodological framework.

The bibliography is extensive and appropriate, incorporating recent literature together with official technical documentation relevant to the evaluated systems. References are properly integrated throughout the manuscript and support both the experimental methodology and the accompanying discussion.

Overall, *Practical Benchmarking and Hands-On Evaluation of AI Popular Systems* represents a valuable scientific contribution to the growing field of AI system evaluation. It successfully bridges the gap between theoretical descriptions of modern AI models and their practical engineering performance, providing readers with experimentally grounded evidence regarding the strengths and limitations of today's leading AI assistants. The work will be particularly valuable for researchers, AI engineers, graduate students, educators, and practitioners interested in objective assessment methodologies for generative artificial intelligence.

Sincerely,

Prof. Dr. Goran Rafajlovski

SRH University of Applied Sciences, Berlin