Chinastandards.org

ICS 35.240
CCS L 70

NATIONAL STANDARD OF THE

PEOPLE'S REPUBLIC OF CHINA

GB/Z 254-2026

Artificial intelligence - Technical requirements for cloud-based inference deployment toolchain of large language models

人工智能 语言大模型云端推理部署工具链技术要求

(English Translation)

Issued on: August 27, 2026 **Implemented on: August 27, 2026**

Issued by: State Administration for Market Regulation;
Standardization Administration of China

Preview: page 1 of 2 shown, 15 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-z-254-2026

Chinastandards.org

Table of Contents

Foreword

1 Scope

2 Normative references

3 Terms and definitions

Foreword

GB/Z 254-2026 is a national guiding technical document (GB/Z) of the People's Republic of China, issued on 27 August 2026 by the State Administration for Market Regulation and the Standardization Administration of China.

It is under the responsibility of the Standardization Administration of China.

Classification: ICS 35.240; CCS L 70.

1 Scope

Running a large language model in the cloud requires a chain of software tools that prepare, optimize and deploy the model for inference, and this guiding document sets technical requirements for that cloud-based inference deployment toolchain. It sits in the artificial intelligence group of GB/Z documents, so it provides guidance and not mandatory obligations.

Toolchain developers and cloud platform providers can use it when designing and describing their products, while enterprises that deploy language models on cloud infrastructure can use it as a reference for selection and for writing procurement requirements. Machine learning engineers, operations teams and independent evaluators who must compare deployment tools will also find that a shared set of requirements gives them common terms. The document was issued on 27 August 2026 by the national standardization administration. It is relevant to suppliers and analysts who follow China's domestic large model ecosystem, in which deployment tooling plays an important role in moving models from research into production services, and to anyone assessing tooling compatibility with national guidance.

Preview: page 2 of 2 shown, 15 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-z-254-2026