Chinastandards.org

ICS 35.240
CCS L70

NATIONAL STANDARD OF THE

PEOPLE'S REPUBLIC OF CHINA

GB/Z 243-2026

Artificial intelligence - Technical requirements for on-device large-scale model inference engine

人工智能 端侧大模型推理引擎技术要求

(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, 17 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-z-243-2026

Chinastandards.org

Table of Contents

Foreword

1 Scope

2 Normative references

3 Terms and definitions

Foreword

GB/Z 243-2026 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 L70.

1 Scope

Running a large model directly on a phone, vehicle or other terminal, instead of in the cloud, depends on a specialized inference engine, and this guidance document covers it. According to its scope, it establishes the reference architecture of on-device large-scale model inference engines from a functional perspective and specifies the technical requirements for deploying large models on devices. It applies to their design, development, testing and deployment.

The requirements are grouped under data, lightweight deployment, adaptation to heterogeneous software and hardware, inference computing acceleration, and inference memory optimization. The document relies on earlier national documents on artificial intelligence terminology and on general requirements for large-scale models.

It is a GB/Z document, so it gives guidance and not mandatory requirements. Chipmakers, terminal and device manufacturers, inference framework developers and test organizations can use it as a design and evaluation reference. Issued on 27 August 2026 by the national standardization administration.

Preview: page 2 of 2 shown, 17 in the full document. Complete English PDF, instant download:
https://chinastandards.org/standards/gb-z-243-2026