ICS 35.020
CCS L 70

中华人民共和国国家标准

GB/T 45280—2025

人工智能　异构人工智能加速器统一接口

Artificial intelligence—Unified interfaces for heterogeneous artificial intelligence accelerating units

2025-02-28 发布　　2025-02-28 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言 …… Ⅲ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　缩略语 …… 2
5　概述 …… 2
5.1　接口功能 …… 2
5.2　架构 …… 2
5.3　基本概念 …… 3
6　接入方法 …… 4
6.1　加速器 …… 4
6.2　机器学习框架 …… 4
6.3　运行过程说明 …… 4
7　接口要求 …… 5
7.1　接口执行状态 …… 5
7.2　接口参数 …… 5
7.3　精度 …… 5
7.4　枚举 …… 5
8　接口定义 …… 6
8.1　计算图表示接口 …… 6
8.2　运行时接口 …… 18
8.3　算子表示接口 …… 44
9　符合性测试方法 …… 46
9.1　通则 …… 46
9.2　测试过程 …… 47
9.3　指标及测量方法 …… 48
附录 A（规范性）　返回码 …… 51
附录 B（规范性）　枚举 …… 54
附录 C（规范性）　算子定义 …… 57
附录 D（资料性）　领域接口 …… 78
附录 E（资料性）　接口示例 …… 83
附录 F（规范性）　测试项 …… 85
参考文献 …… 87

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、华为技术有限公司、上海人工智能创新中心、北京航空航天大学、上海燧原科技股份有限公司、北京壁仞科技开发有限公司、上海天数智芯半导体有限公司、中科寒武纪科技股份有限公司、英特尔(中国)有限公司、深圳云天励飞技术股份有限公司、沐曦集成电路(上海)有限公司、中国科学院软件研究所、浪潮电子信息产业股份有限公司、北京大学、上海市人工智能行业协会、华为云计算技术有限公司、上海商汤智能科技有限公司、北京智芯微电子科技有限公司、平头哥(上海)半导体技术有限公司、杭州海康威视数字技术股份有限公司、中国移动通信集团有限公司、深圳鲲云信息科技有限公司、龙芯中科技术股份有限公司、南京南瑞瑞腾科技有限责任公司、苏州登临科技有限公司、南方电网人工智能科技有限公司、美的集团(上海)有限公司、北京大学长沙计算与数字经济研究院、北京航空航天大学杭州创新研究院、四川华鲲振宇智能科技有限责任公司、山东浪潮科学研究院有限公司、中国南方电网有限责任公司、西南科技大学、浙江大华技术股份有限公司、北京格灵深瞳信息技术股份有限公司、北京电子数智科技有限责任公司。

本文件主要起草人：董建、杨雨泽、张亚丽、徐洋、张行程、鲍薇、刘文枫、裴芝林、王莞尔、曹晓琦、马骋昊、栾钟治、梅敬青、丁瑞全、胡铭珊、程归鹏、王海宁、苏岚、刘梓、孟令中、马珊珊、李斌斌、宿栋栋、杨超、赵春昊、刘勇、钟普、张艺伯、章放、金镝、蔡权雄、马莞悦、石超、慈红斌、陈柔伊、蔡亚森、贾梦珠、胡征慧、赵彦钧、李锐、张喜铭、徐欢、俞文心、方贵明、于杰、郭文。

人工智能　异构人工智能加速器统一接口

1　范围

本文件定义了异构人工智能加速器的统一接口及其语义和使用方法，描述了各加速器为实现此接口所需的接入方法和试验方法。

本文件适用于人工智能加速器接口的设计和实现，也可为人工智能加速器应用提供参考。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867—2022　信息技术　人工智能　术语

GB/T 45087—2024　人工智能　服务器系统性能测试方法

GB/T 17966—2024　信息技术　微处理器系统　浮点运算

3　术语和定义

GB/T 41867－2022 界定的以及下列术语和定义适用于本文件。

3.1

[应用编程]接口 [application programming] interface

〈人工智能〉用来使用人工智能加速器功能的语法和语义定义。

[来源：ISO/IEC/IEEE 9945:2009，3.19，有修改]

3.2

人工智能加速[处理]器　artificial intelligence accelerating processor

人工智能加速芯片　artificial intelligence accelerating chip

具备适配人工智能算法的运算微架构，能够完成人工智能应用运算处理的集成电路元件。

注：在本文件中，在不引起误解的语境中，将人工智能加速器简称为加速器。

[来源：GB/T 41867—2022，3.1.5]

3.3

计算图　computational graph

用来表示数学函数，由节点和连接构成的有向图。

注 1：节点表示数学运算，即算子。

注 2：连接表示数学运算之间的依赖关系。

注 3：一个连接联通起始节点和终止节点。

[来源：ISO/IEC/IEEE 24765:2017，3.1762.1，有修改]