

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2026 年 3 月 5 日 (05.03.2026)



(10) 国际公布号
WO 2026/045029 A1

(51) 国际专利分类号:

G06F 16/332 (2025.01) **G06N 20/20** (2019.01)
G06F 40/289 (2020.01) **G06N 3/045** (2023.01)
G06F 18/214 (2023.01) **G06N 3/08** (2023.01)

(21) 国际申请号: PCT/CN2024/140678

(22) 国际申请日: 2024 年 12 月 19 日 (19.12.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:

202411191438.2 2024 年 8 月 28 日 (28.08.2024) CN

(71) 申请人: 无限光年 (上海) 技术有限公司 (INFLY TECH (SHANGHAI) CO., LTD.) [CN/CN]; 中国上海市徐汇区虹漕路 25-1 号 2 层 1325 室 200233 (CN)。

(72) 发明人: 褚巍 (CHU, Wei); 中国上海市徐汇区虹漕路 25-1 号 2 层 1325 室 200233 (CN)。

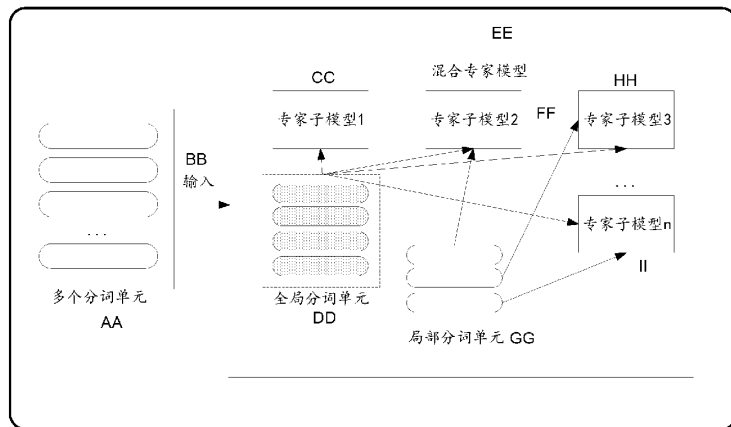
(74) 代理人: 北京智信禾专利代理有限公司 (CHISPO ATTORNEYS AT LAW); 中国北京市东城区崇文门外大街 8 号院哈德门广场西塔 901 100062 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,

(54) **Title:** MODEL OPTIMIZATION METHOD AND APPARATUS, AND INFORMATION PROCESSING METHOD AND APPARATUS

(54) 发明名称: 模型优化方法及装置、信息处理方法及装置



AA A plurality of tokens
BB Input
CC Expert sub-model
DD Global tokens
EE Mixture-of-experts model
FF Local tokens

图 1

(57) **Abstract:** Provided are a model optimization method and apparatus, and an information processing method and apparatus. The model optimization method comprises: inputting a plurality of tokens into a mixture-of-experts model, and by means of a router in the mixture-of-experts model, calculating a target entropy value corresponding to each token; on the basis of the target entropy values, dividing the plurality of tokens into global tokens and local tokens; broadcasting the global tokens to all expert sub-models in the mixture-of-experts model for processing, and broadcasting the local tokens to expert sub-models, which match the local tokens, in the mixture-of-experts model for processing; and on the basis of processing results, adjusting parameters of the mixture-of-experts model until a target mixture-of-experts model which meets an optimization stop condition is obtained.

[见续页]

HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO,
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,
CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN,
TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 提供了一种模型优化方法及装置、信息处理方法及装置, 其中所述模型优化方法包括: 将多个分词单元输入至混合专家模型, 通过所述混合专家模型中的路由器计算每个分词单元对应的目标熵值; 根据所述目标熵值将所述多个分词单元划分为全局分词单元和局部分词单元; 将所述全局分词单元广播至所述混合专家模型中的全部专家子模型进行处理, 以及将所述局部分词单元广播至所述混合专家模型中匹配所述局部分词单元的专家子模型进行处理; 根据处理结果对所述混合专家模型进行调参, 直至获得满足优化停止条件的目标混合专家模型。