

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
04 April 2024 (04.04.2024)



(10) International Publication Number
WO 2024/068784 A1

(51) International Patent Classification:

G06F 16/332 (2019.01) G06N 3/02 (2006.01)
G06F 16/33 (2019.01) G06N 3/092 (2023.01)
G06F 16/338 (2019.01) G06N 5/04 (2023.01)

(21) International Application Number:

PCT/EP2023/076792

(22) International Filing Date:

27 September 2023 (27.09.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/377,532 28 September 2022 (28.09.2022) US

(71) Applicant: DEEPMIND TECHNOLOGIES LIMITED

[GB/GB]; 5 New Street Square, London EC4A 3TW (GB).

(72) Inventors: HIGGINS, Irina; 6 Pancras Square, London

NIC 4AG (GB). UESATO, Jonathan Ken; 6 Pancras

Square, London NIC 4AG (GB). KUSHMAN, Nathaniel

Arthur; 6 Pancras Square, London NIC 4AG (GB). KU-

MAR, Ramana; 6 Pancras Square, London NIC 4AG

(GB).

(74) Agent: FISH & RICHARDSON P.C.; Highlight Business
Towers, Mies-van-der-Rohe-Str. 8, 80807 Munich (DE).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available):

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) Designated States (unless otherwise indicated, for every
kind of regional protection available):

ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, 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).

(54) Title: REWARD-MODEL BASED REINFORCEMENT LEARNING FOR PERFORMING REASONING TASKS

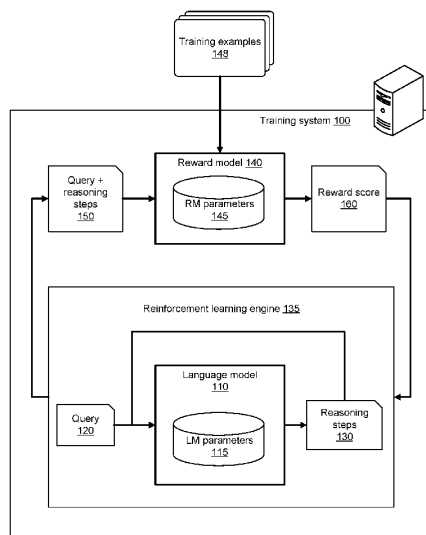


FIG. 1A

(57) Abstract: Methods, systems, and apparatus, including computer programs encoded on computer storage media, for A training a language model for performing a reasoning task. The system obtains a plurality of training examples. Each training example includes a respective sample query text sequence characterizing a respective sample query and a respective reference response text sequence that includes a reference final answer to the respective sample query. The system trains a reward model on the plurality of training examples. The reward model is configured to receive an input including a query text sequence characterizing a query and one or more reasoning steps that have been generated in response to the query and process the input to compute a reward score indicating how successful the one or more reasoning steps are in yielding a correct final answer to the query. The system trains the language model using the trained reward model.

WO 2024/068784 A1

[Continued on next page]

Published:

— with international search report (Art. 21(3))