



(51) International Patent Classification:

G06N 3/045 (2023.01) G06N 3/047 (2023.01)
G06F 40/284 (2020.01)

(21) International Application Number:

PCT/US2024/017336

(22) International Filing Date:

26 February 2024 (26.02.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/460,850 20 April 2023 (20.04.2023) US
18/538,912 13 December 2023 (13.12.2023) US

(71) Applicant: **QUALCOMM INCORPORATED** [US/US];

ATTN: International IP Administration, 5775 Morehouse
Drive, San Diego, California 92121-1714 (US).

(72) Inventors: **LOTT, Christopher**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **LEE, Mingu**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **JEON, Wonseok**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **MEMISEVIC, Roland**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US).

(74) Agent: **ROBERTS, Steven E.** et al.; Patterson + Sheridan, L.L.P., 24 Greenway Plaza, Suite 1600, Houston, Texas 77046-2472 (US).

(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,

(54) Title: SPECULATIVE DECODING IN AUTOREGRESSIVE GENERATIVE ARTIFICIAL INTELLIGENCE MODELS

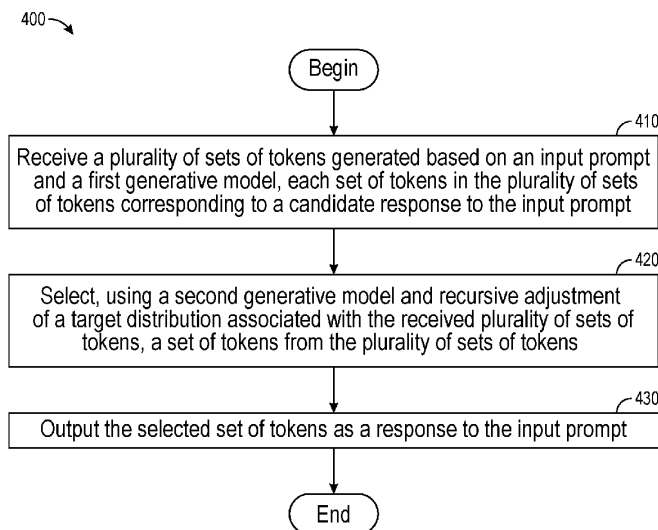


FIG. 4

(57) Abstract: Certain aspects of the present disclosure provide techniques and apparatus for generating a response to a query input in a generative artificial intelligence model. An example method generally includes receiving a plurality of sets of tokens generated based on an input prompt and a first generative artificial intelligence model, each set of tokens in the plurality of sets of tokens corresponding to a candidate response to the input prompt; selecting, using a second generative artificial intelligence model and recursive adjustment of a target distribution associated with the received plurality of sets of tokens, a set of tokens from the plurality of sets of tokens; and outputting the selected set of tokens as a response to the input prompt.

[Continued on next page]

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).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*