



(51) International Patent Classification:
G06F 40/242 (2020.01) *H04L 9/40* (2022.01)
G06F 21/55 (2013.01)

(21) International Application Number:
PCT/US2025/013797

(22) International Filing Date:
30 January 2025 (30.01.2025)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
18/429,134 31 January 2024 (31.01.2024) US
18/429,255 31 January 2024 (31.01.2024) US

(71) Applicant: **HIDDENLAYER, INC.** [US/US]; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).

(72) Inventors: **CAPPEL, Kwesi**; c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **BURNS, Tanner**; c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **YEUNG, Kenneth**; c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).

(74) Agent: **KUKKONEN, Carl, A., III** et al.; JONES DAY, 250 Vesey Street, New York, NY 10281-1047 (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, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UY, 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: GENERATIVE ARTIFICIAL INTELLIGENCE MODEL PROTECTION USING OUTPUT BLOCKLIST

(57) Abstract: The inputs and/or outputs of a generative artificial intelligence model are monitored to determine whether they contain or otherwise elicit undesired behavior by the model such as bypassing security measures, leaking sensitive information, or generating or consuming malicious content. This determination can be used to selectively trigger remediation processes to protect the model from malicious actions. Related apparatus, systems, techniques and articles are also described.

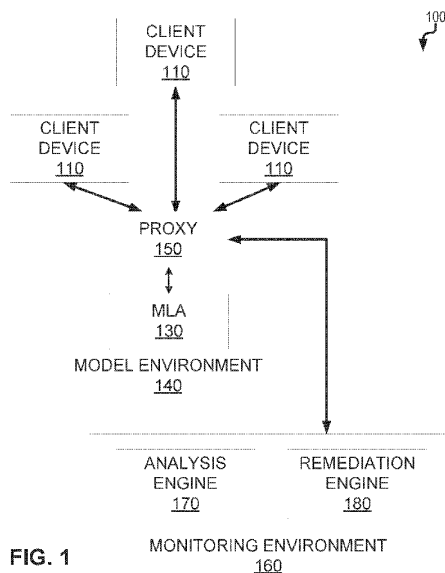


FIG. 1

Published:

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