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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2025 (18.09.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/193532 A1

(51) International Patent Classification:
G06F 21/55 (2013.01) *G06F 21/57* (2013.01)
H04L 9/40 (2022.01) *G06N 20/00* (2019.01)

(21) International Application Number:
PCT/US2025/018863

(22) International Filing Date:
07 March 2025 (07.03.2025)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
18/605,337 14 March 2024 (14.03.2024) US
19/038,505 27 January 2025 (27.01.2025) US

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

(72) Inventors: **RIDEOUT, Jacob**; 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). **CAPPEL, Kwesi**; 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.** 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,

(54) Title: GENERATIVE ARTIFICIAL INTELLIGENCE MODEL PROMPT INJECTION CLASSIFIER

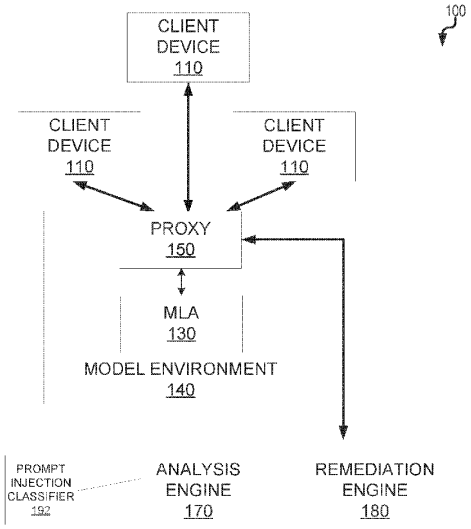


FIG. 1

(57) Abstract: An analysis engine receives data characterizing a prompt for ingestion by a generative artificial intelligence (GenAI) model. The analysis engine, using a prompt injection classifier determines whether the prompt comprises or is indicative of malicious content or otherwise elicits malicious actions. Data characterizing the determination is provided to a consuming application or process. Related apparatus, systems, techniques and articles are also described.

[Continued on next page]

WO 2025/193532 A1

SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*