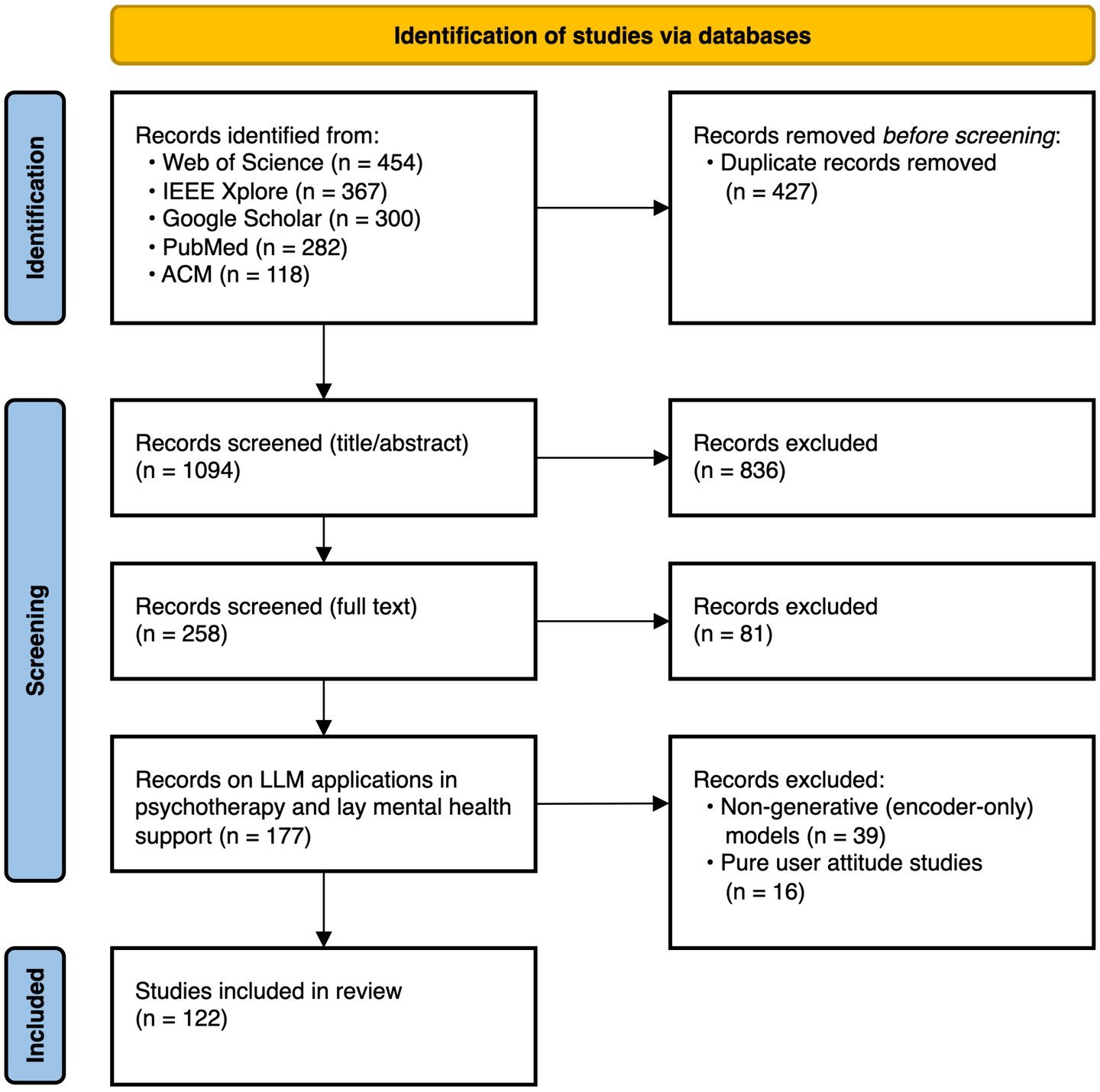

Figure S1 | PRISMA 2020 flow diagram for the study selection process. Shown are the number of records identified from various databases, the number of records screened, and the final number of studies included in the review, along with the reasons for exclusion at each stage. *Abbreviations: ACM, Association for Computing Machinery; IEEE, Institute of Electrical and Electronics Engineers; LLM, Large Language Model; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.*

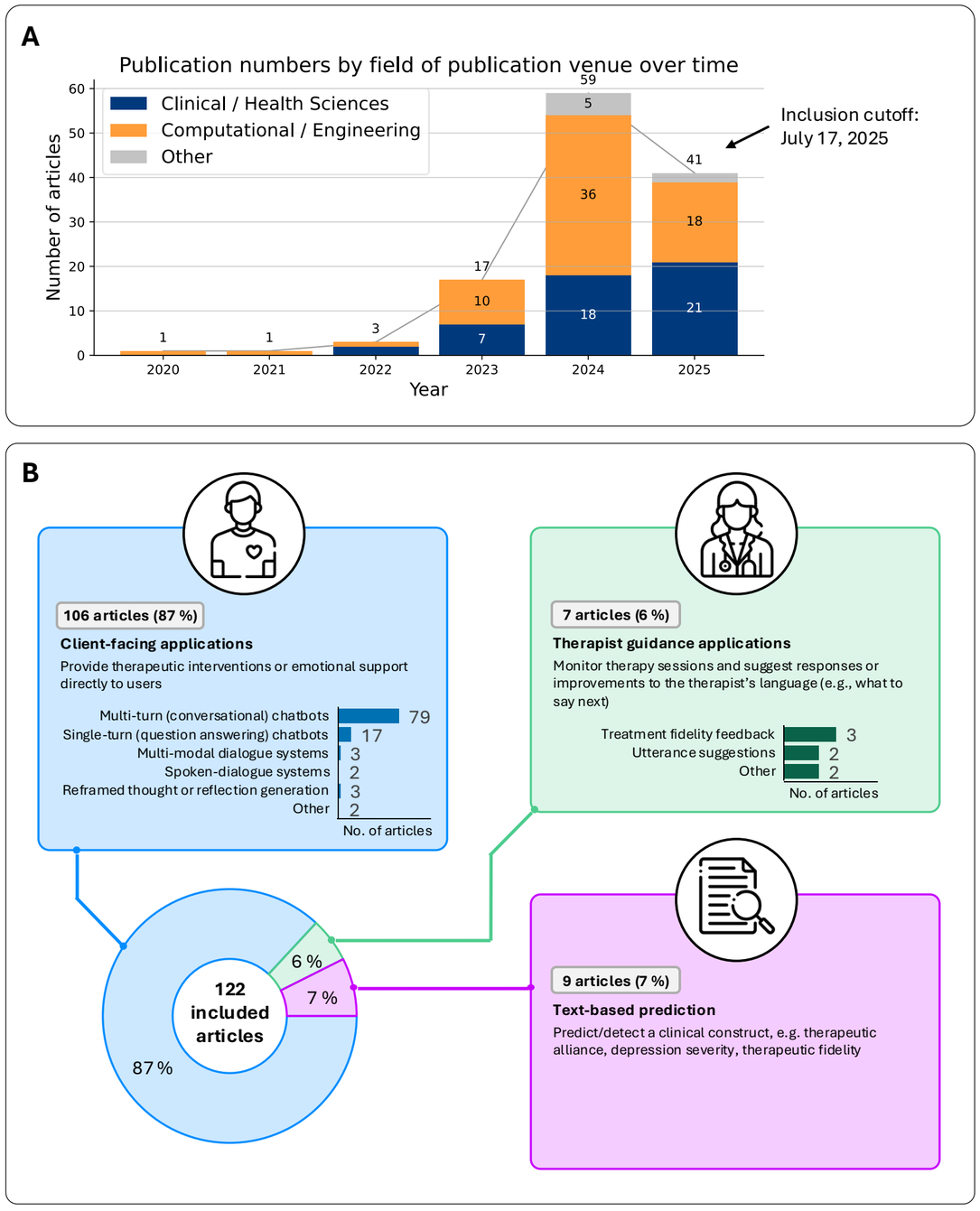

Figure 1 | Publication trends and application types in included articles. The figure summarizes the 122 articles included in the review by application type and publication trend. **A**, Number of articles published annually from 2020 to 2025. The stacked bar chart shows the total count per year, categorized by the field of the publication venue (i.e., journal or conference) (Clinical/Health Sciences, Computational/Engineering, or Other). The data for 2025 reflects a partial year, with an inclusion cutoff date of July 17, 2025. **B**, Classification of the articles into three primary LLM application types. The donut chart illustrates the overall distribution, with detailed cards providing a further breakdown of sub-types and their respective article counts for client-facing applications, therapist guidance applications, and text-based prediction.

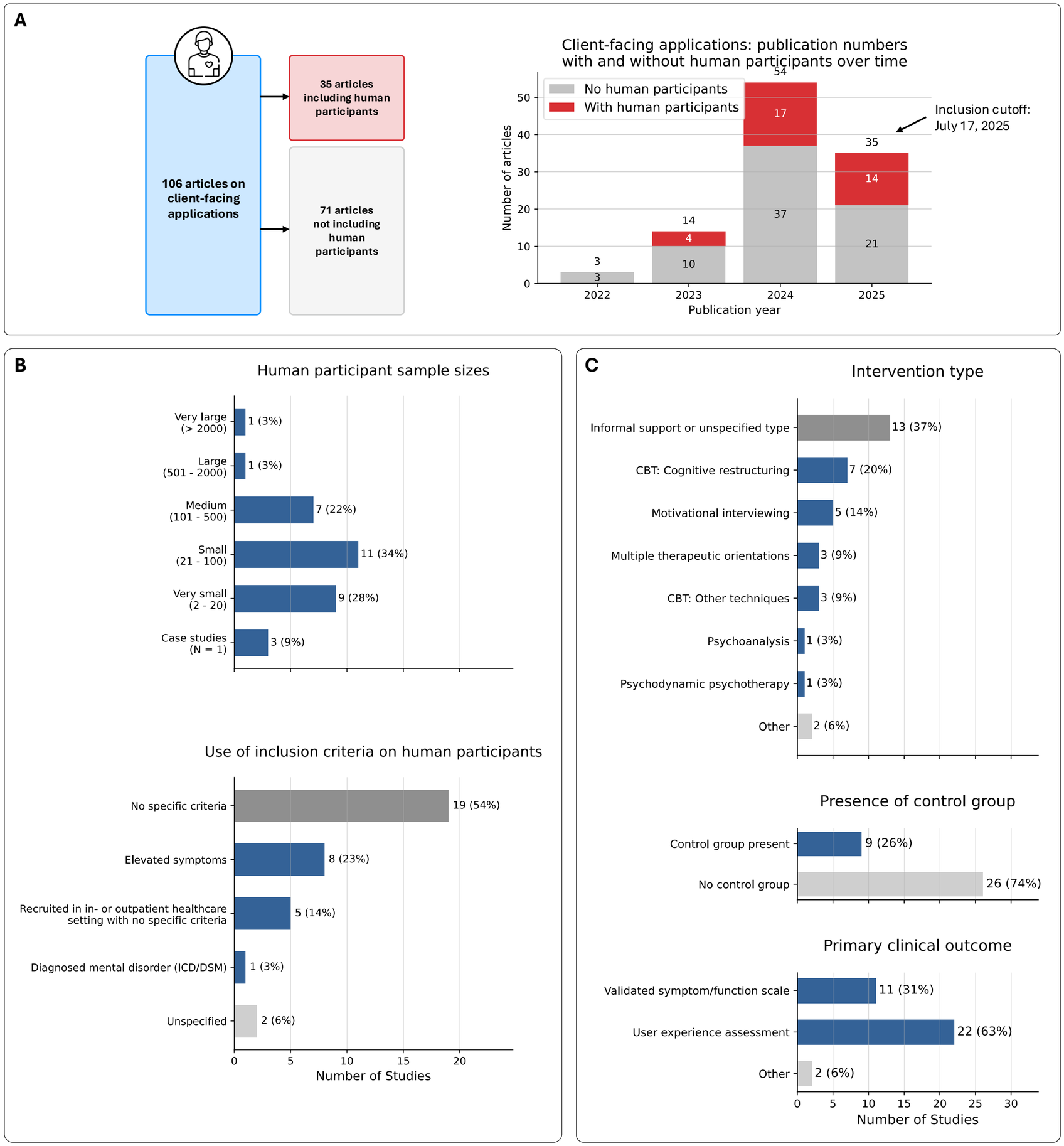

Figure 2 | Methodological characteristics of the 35 client-facing application studies involving human participants. A, The flowchart illustrates the proportion of articles on client-facing applications that included human participants (35 of 106). The stacked bar chart shows the annual publication numbers from 2022 to 2025, distinguishing between studies with human participants (red) and without (grey). The data for 2025 reflects a partial year due to an inclusion cutoff of July 17, 2025. **B**, Bar charts showing the distribution of the 35 studies across different sample size categories (top) and by the type of inclusion criteria used (bottom). **C**, Bar charts classifying the 35 studies by their primary intervention type, the presence of a control group, and the nature of the primary clinical outcome. *Abbreviations: CBT, Cognitive Behavioral Therapy; ICD/DSM, International Classification of Diseases/Diagnostic and Statistical Manual of Mental Disorders.*

Study	Population	Intervention	Control Condition	Group Allocation	Primary Outcome(s)	Between-group Effect Size (d)
Heinz et al. (2025)	Adults with clinically significant symptoms of MDD, GAD, or at high risk for eating disorders (n = 210)	4-week intervention with a fine-tuned generative AI chatbot (Therabot)	Waitlist control	Randomized	Change in depression (PHQ-9) and anxiety (GAD-Q-IV) symptoms, and weight concerns (WCS)	0.63-0.90
Habicht et al. (2025)	Adult patients in UK's NHS services (n = 244)	GPT-4-based therapy support tool (Limbic Care) augmenting human-led 6-session group CBT with between-session exercises	Standard care (static CBT worksheets) within the same group-CBT context	Self-selection by participants	Change in depression (PHQ-9) and anxiety (GAD-7) symptoms, measured as treatment success rates	≈0.44–0.57*
Kolenik et al. (2024)	Non-clinical convenience sample (n = 42)	Single session with a self-developed GPT-3-based chatbot	Single session with rule-based chatbot (Woebot)	Randomized	Change in stress, anxiety, and depression symptoms (SISQs)	Not reported
Melo et al. (2024)	Psychiatric inpatients admitted for suicidal ideation (n = 12)	3-6 guided sessions with ChatGPT (version GPT-3.5)	Standard inpatient psychiatric care (treatment as usual)	Randomized	Change in quality of life (WHOQOL-BREF)	≈1.56*
Liu et al. (2024)	Non-clinical, China-based online participants (n = 48)	2-week intervention; GPT-3.5 provided real-time feedback on exercises within a rule-based positive psychology intervention chatbot	Active control; same rule-based chatbot intervention but without real-time feedback	Randomized	Change in depression (PHQ-9) and anxiety (GAD-7) symptoms, and life satisfaction (SWLS)	0.01-0.59
Nazarova (2023)	Non-clinical university students (n = 68)	8-week intervention with a GPT-3 based CBT chatbot (TeaBot)	No intervention; access to a psychoeducation manual only	Randomized	Change in cognitive distortions and psychological flexibility (AAQ + CDS sum score)	Not reported

Table 1 | Overview of controlled clinical studies on client-facing LLM applications. The table summarizes studies included in this review, selected based on three criteria: (1) they evaluated a client-facing application involving human participants, (2) they included a control or comparison group, and (3) they reported outcomes using validated clinical symptom or function scales. For each study, the population, intervention, control condition, group allocation method, primary outcomes, and the between-group effect size (Cohen's *d*) are presented.
These effect sizes are approximated based on reported statistics (e.g., converted from Odds Ratios) as Cohen's d was not directly provided in the text.* ***Abbreviations**: AAQ, Acceptance and Action Questionnaire; CBT, Cognitive Behavioral Therapy; CDS, Cognitive Distortions Scale; CHR-FED, Clinically High Risk for Feeding and Eating Disorders; GAD, Generalized Anxiety Disorder; GAD-Q-IV, GAD Questionnaire; MDD, Major Depressive Disorder; NHS, National Health Service; PHQ-9, Patient Health Questionnaire-9; PPI, Positive Psychology Intervention; SISQs, Single Item Screening Questions; SWLS, Satisfaction With Life Scale; WCS, Weight Concerns Scale; WHOQOL-BREF, World Health Organization Quality of Life-Brief Version.*

- Heinz, Michael V., et al. "Randomized trial of a generative AI chatbot for mental health treatment." *Nejm Ai* 2.4 (2025): AIoa2400802.
- Habicht, Johanna, et al. "Generative AI–enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study." *Journal of medical Internet research* 27 (2025): e60435.
- Kolenik, Tine, Günter Schiepek, and Matjaž Gams. "Computational psychotherapy system for mental health prediction and behavior change with a conversational agent." *Neuropsychiatric Disease and Treatment* (2024): 2465-2498.
- Melo, Antonio, Inês Silva, and Joana Lopes. "Chatgpt: A pilot study on a promising tool for mental health support in psychiatric inpatient care." *International Journal of Psychiatric Trainees* 2.2 (2024).
- Liu, Ivan, et al. "Investigating the key success factors of chatbot-based positive psychology intervention with retrieval-and generative pre-trained transformer (GPT)-based chatbots." *International Journal of Human–Computer Interaction* 41.1 (2025): 341-352.
- Nazarova, Deniz. "Application of artificial intelligence in mental healthcare: generative pre-trained transformer 3 (Gpt-3) and cognitive distortions." *Proceedings of the Future Technologies Conference*. Cham: Springer Nature Switzerland, 2023.

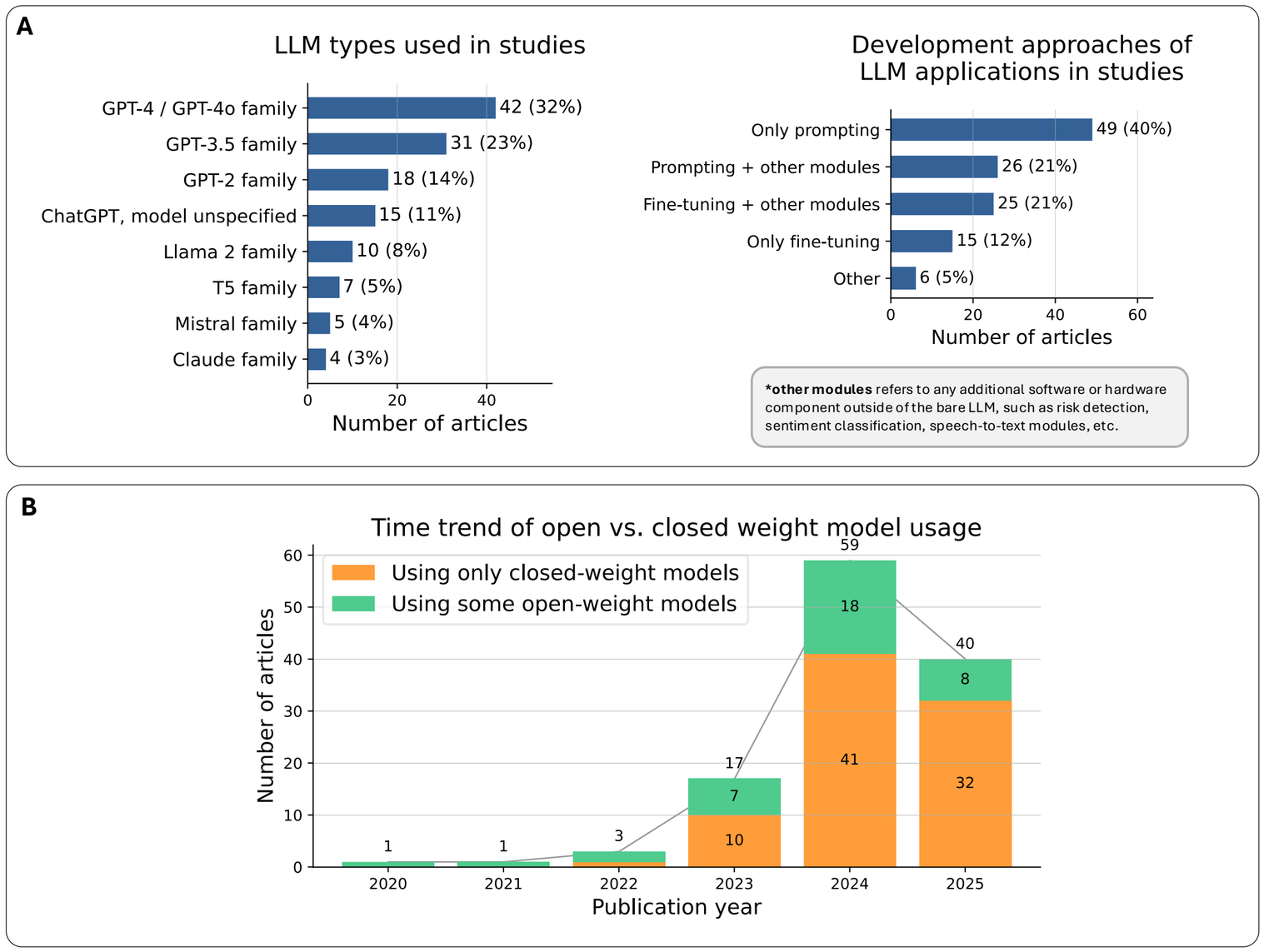

Figure 3 | Technical characteristics of the reviewed LLM applications. A, Horizontal bar charts detailing technical characteristics: the frequency of the top 8 Large Language Model (LLM) families utilized (left; note that a single study may utilize multiple LLMs and thus be counted in multiple categories), and the distribution of development approaches (right). **B**, A stacked bar chart illustrating the annual publication trend from 2020 to 2025, categorized by model weight accessibility. The chart distinguishes between applications built using exclusively closed-weight models (where model weights are proprietary) and those incorporating at least some open-weight models (where model weights are publicly available for local deployment).

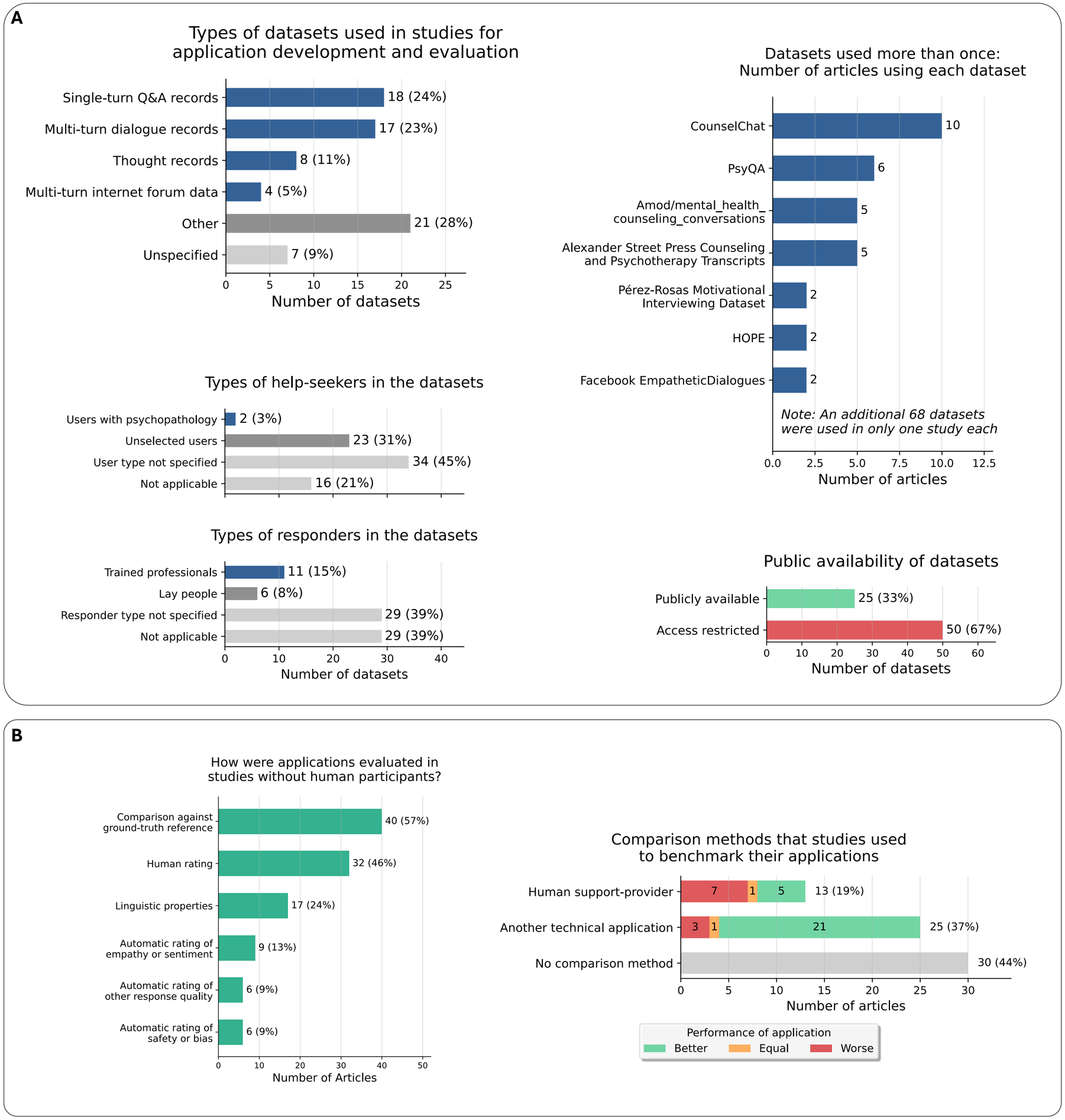

Figure 4 | Characteristics of datasets and evaluation methods. This figure details the types of datasets used in the reviewed studies and how applications were evaluated. **A**, The top left bar chart categorizes datasets by their type, such as single-turn Q&A or multi-turn dialogue records. Below it, the bar charts show the characteristics of help-seekers (middle left) and responders (bottom left) within these datasets. The bar chart on the right identifies specific datasets that were used in more than one article, indicating their reuse across studies. **B**, This panel focuses on how applications were evaluated in studies without human participants. The left bar chart shows the frequency of different evaluation approaches (e.g., comparison against ground-truth reference, human rating, linguistic properties). The right stacked bar chart categorizes comparison methods (e.g., human support-provider, another technical application, no comparison) and, for those with comparisons, indicates the reported performance of the application (better, equal, or worse).

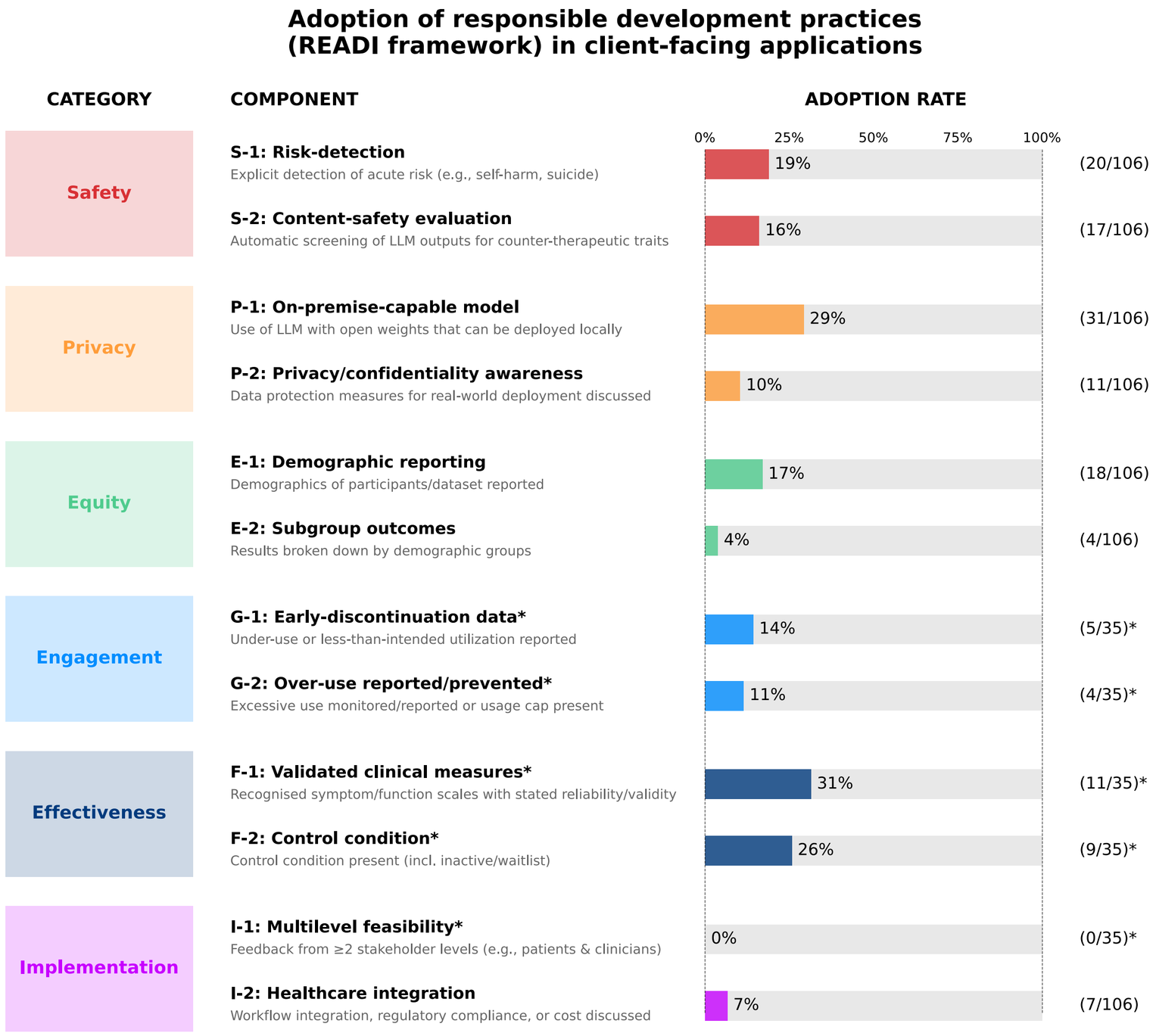

Figure 5 | Adoption of responsible development practices (READI framework) in client-facing applications. The figure displays the adoption rate of 12 responsible development practices, organized into the six categories of the READI framework (Safety, Privacy, Equity, Engagement, Effectiveness, and Implementation). Each bar indicates the percentage of articles that adopted a specific component, with the raw counts shown on the right.
**The denominator for most components is the total number of client-facing applications (n = 106). Components marked with an asterisk were assessed only in the subset of studies that involved human participants (n = 35).*
***Abbreviations**: LLM, Large Language Model.*

- Stade, Elizabeth C., et al. "Readiness evaluation for artificial intelligence-mental health deployment and implementation (READI): a review and proposed framework." (2025).