Supplementary Information

A Self-Evolving Agent for Longitudinal Personal Health Management

HealthJev

Supplementary Tables S1–S12; Figures S1–S6.

Supplementary Table S1. Source resources and access conditions

Resources support the record-grounded longitudinal benchmark and the six biomedical tasks. Obtain each resource under its provider terms. MIMIC-IV requires credentialed access and a data-use agreement; raw records and images are not redistributed.

Resource	Evaluation role	Access conditions	Provider
MIMIC-IV	Longitudinal reference trajectories; MIMIC-ICD coding	PhysioNet credentialed access and data-use agreement	https://physionet.org/content/mimiciv/
Diabetes 130-US hospitals	Three-class readmission prediction	Obtain under UCI provider terms	https://doi.org/10.24432/C5230J
GeneTuring	Genomic question answering	Obtain the benchmark under the original provider terms	https://doi.org/10.1093/bib/bbaf492
ODIR-5K	Eight-class fundus diagnosis	Obtain under the challenge provider terms	https://odir2019.grand-challenge.org/
PhysioNet/CinC Challenge 2012	ICU SOFA estimation	Obtain under the PhysioNet resource terms	https://physionet.org/content/challenge-2012/
MedCalc-Bench	Medical calculation	Obtain under the original benchmark provider terms	https://github.com/ncbi-nlp/MedCalc-Bench

Supplementary Table S2. Anonymous users and simulated interaction personas

The 20 anonymous user labels match main Fig. 2b. Ages range from 22 to 74 years. Occupations and persona tags specify simulated interaction contexts, not patient identities or psychological measurements. The same persona is used throughout each 90-day interaction. This table does not reproduce clinical intake profiles or source identifiers.

User	Age (years)	Simulated occupation	Simulated persona tag
U01	48	Rare-book proofreader	dutiful
U02	57	Retired primary-school art teacher	dutiful
U03	50	Retired community-library administrator	steady
U04	60	Retired primary-school Chinese-language teacher	dutiful
U05	22	Content-operations intern	anxious
U06	49	State-owned-enterprise warehouse administrator	anxious
U07	34	Backend developer	steady
U08	49	Retired community librarian	steady
U09	57	Retired cultural-centre folk-dance instructor	laid-back
U10	46	Neighbourhood grocery-store owner	terse
U11	62	Retired kindergarten childcare worker	terse
U12	55	Retired horticultural technician	laid-back
U13	65	Retired municipal arts-centre dance instructor	laid-back
U14	68	Retired symphony-orchestra violinist	chatty
U15	65	Retired textile-machine operator	chatty
U16	45	Secondary-school cashier	avoidant
U17	31	Landscape draftsperson	self-experimenter
U18	74	Retired makeup artist	self-experimenter
U19	71	Retired secondary-school cafeteria pastry cook	self-experimenter
U20	61	Retired postal newspaper-distribution worker	avoidant

Supplementary Table S3. Longitudinal benchmark composition

The benchmark comprises 20 users, 90 days per user and 4,680 goals across seven categories.

Category	Item	Count / range	Unit
Scale	Users	20	users
Scale	Interaction days	1800	user-days
Scale	Daily goals	4680	goals
Goal composition	Lifestyle consult	1332	goals
Goal composition	Event report	1161	goals
Goal composition	Daily chat	1080	goals
Goal composition	Help seeking	579	goals
Goal composition	Self-monitoring	264	goals
Goal composition	Latest-value / supersession	180	goals
Goal composition	Medication consult	84	goals
Historical context	Historical references	1696	references
Historical context	Days with historical references	1544	user-days
Historical context	Reference lag	1–25	days
Privacy	Privacy-scenario days	720	user-days
Governance	Governance probes	189	probes
Governance	Governance checks	294	checks

Supplementary Table S4. Six biomedical test tasks and scoring rules

Test sets and task-specific scores correspond to main Fig. 3a. The reported primary metric is the mean case score. GeneTuring, ICU SOFA and MIMIC-ICD preserve the partial-credit or set-based scoring rules specified in the main Methods.

Task	Test n	Input	Output	Primary case score
Diabetes readmission	200	Electronic health record	Three classes: NO, <30 days, >30 days	Exact class match
GeneTuring	150	Genomic question	Numeric answer, gene symbol or sequence	Numeric agreement; normalized short-answer matching; SequenceMatcher similarity for long sequences
ODIR5K	203	Fundus image	One of eight diagnostic classes	Exact class match
ICU SOFA	200	Physiological measurements	Total SOFA score	Exact score: 1; absolute error 1: 0.5; otherwise: 0
MIMIC-ICD	201	Discharge summary	Subset of the 30 candidate ICD-10 codes	Set F1 against reference codes
MedCalc	200	Clinical calculation question	Value, date or text	Programmatic value, date and text matching

Supplementary Table S5. Evaluation conditions and task evidence tools

The longitudinal experiment compares five conditions; biomedical testing additionally includes Base. Base answers without tools or memory. The scaffold-mounted conditions share task evidence tools and scoring. Biomedical experience libraries are frozen and writeback is disabled during testing. Reference answers enter review after submission, not tool inputs.

Category	Condition or task	Protocol or tool
Condition	Base	No tools or memory; direct task response
Condition	ReAct	Tool-augmented scaffold without persistent experience memory
Condition	Mem0	Same scaffold with Mem0 memory
Condition	Memp	Same scaffold with procedural experience memory
Condition	MemoryBank	Same scaffold with MemoryBank memory
Condition	HealthJev	Same scaffold with screened and utility-weighted experience
Task tools	Diabetes readmission	Literature search and statistical readmission-prior lookup
Task tools	GeneTuring	Gene-database entry point: MyGene, Ensembl, HGNC, NCBI E-utilities and Open Targets
Task tools	ODIR5K	No task-specific tool; direct multimodal interpretation
Task tools	ICU SOFA	No task-specific tool
Task tools	MIMIC-ICD	Candidate ICD-10 dictionary lookup
Task tools	MedCalc	Structured calculator covering 55 reference formulas

Supplementary Table S6. Final longitudinal daily scores across five conditions

HealthJev achieves the highest mean score. Each condition includes 20 users × 90 days. Confidence intervals resample users as whole clusters (10,000 resamples; seed 20260705). Two-sided paired Wilcoxon tests compare daily scores against HealthJev; the pairing unit differs from the confidence-interval resampling unit.

Condition	**Paired user-days**	**Mean score**	**95% user-cluster CI**	**p vs HealthJev**
HealthJev	1800	0.667572	0.648444–0.687173	—
MemoryBank	1800	0.633922	0.602031–0.661552	0.000111
Mem0	1800	0.592839	0.569633–0.615790	2.74e-25
ReAct	1800	0.509917	0.496177–0.523389	1.94e-111
Memp	1800	0.501806	0.485339–0.518789	2.59e-124

Supplementary Table S7. Complete six-task, six-condition primary-metric matrix

HealthJev has the highest primary metric in each of the six tasks. Scores are percentages of the mean task-specific case score; the six-task mean gain over Base is 13.3 percentage points. Gains are computed before rounding. Uncorrected two-sided exact McNemar p values compare paired positive-score outcomes (case score > 0). For ICD this tests non-zero-score outcomes, not the mean set F1.

Task	n	Base	ReAct	Mem0	Memp	MemoryBank	HealthJev	Gain (pp)	p vs Base
Diabetes readmission	200	24.50	52.50	53.50	50.50	52.00	55.50	+31.00	1.02e-09
GeneTuring	150	18.71	27.66	33.89	31.13	30.57	35.17	+16.47	4.17e-07
ODIR5K	203	19.70	21.67	22.17	26.11	17.24	35.96	+16.26	1.02e-08
ICU SOFA	200	37.00	33.00	35.25	36.50	38.25	43.00	+6.00	0.125
MIMIC-ICD	201	48.97	46.21	50.03	49.46	49.42	50.14	+1.17	0.008
MedCalc	200	68.00	67.00	69.50	74.00	71.50	77.00	+9.00	0.022

Supplementary Table S8. Frozen-library transfer across seven backbones

The frozen library increases the macro-average on all seven backbones, including six additional backbones. All 42 task-level pairs are reported. Each model is compared with its own Base. Task p values use exact McNemar tests on positive-score outcomes; macro p values use exact Wilcoxon signed-rank tests over six paired task means, enumerating all sign assignments with average ranks for ties. Macro means and differences are calculated before rounding.

Backbone	Task	n / unit	Base %	Library %	Difference pp	p
Doubao-Seed-2.0-pro	Diabetes readmission	200	24.50	55.50	+31.00	1.02e-09
Doubao-Seed-2.0-pro	GeneTuring	150	18.71	35.17	+16.47	4.17e-07
Doubao-Seed-2.0-pro	ODIR5K	203	19.70	35.96	+16.26	1.02e-08
Doubao-Seed-2.0-pro	ICU SOFA	200	37.00	43.00	+6.00	0.125
Doubao-Seed-2.0-pro	MIMIC-ICD	201	48.97	50.14	+1.17	0.008
Doubao-Seed-2.0-pro	MedCalc	200	68.00	77.00	+9.00	0.022
Doubao-Seed-2.0-pro	Macro-average (six tasks)	6 tasks	36.15	49.46	+13.32	0.031
DeepSeek-V4.1-Flash	Diabetes readmission	200	19.00	45.00	+26.00	1.81e-07
DeepSeek-V4.1-Flash	GeneTuring	150	23.91	31.76	+7.85	0.009
DeepSeek-V4.1-Flash	ODIR5K	203	25.12	42.86	+17.73	4.04e-07
DeepSeek-V4.1-Flash	ICU SOFA	200	38.75	42.25	+3.50	0.401
DeepSeek-V4.1-Flash	MIMIC-ICD	201	41.21	51.87	+10.66	6.88e-05
DeepSeek-V4.1-Flash	MedCalc	200	70.50	82.50	+12.00	0.001
DeepSeek-V4.1-Flash	Macro-average (six tasks)	6 tasks	36.42	49.37	+12.96	0.031
GPT-5.4	Diabetes readmission	200	40.00	51.50	+11.50	0.011
GPT-5.4	GeneTuring	150	19.47	29.52	+10.05	0.000221
GPT-5.4	ODIR5K	203	17.73	29.56	+11.82	8.43e-06
GPT-5.4	ICU SOFA	200	38.75	40.50	+1.75	1.000
GPT-5.4	MIMIC-ICD	201	54.57	54.44	-0.12	0.454
GPT-5.4	MedCalc	200	54.50	75.50	+21.00	5.71e-08
GPT-5.4	Macro-average (six tasks)	6 tasks	37.50	46.84	+9.33	0.062
GPT-4.1	Diabetes readmission	200	19.50	44.50	+25.00	2.26e-07
GPT-4.1	GeneTuring	150	16.75	24.06	+7.31	0.007
GPT-4.1	ODIR5K	203	32.51	32.02	-0.49	1.000
GPT-4.1	ICU SOFA	200	28.25	41.75	+13.50	6.49e-05
GPT-4.1	MIMIC-ICD	201	50.08	42.08	-8.01	3.47e-06
GPT-4.1	MedCalc	200	35.00	48.50	+13.50	0.002
GPT-4.1	Macro-average (six tasks)	6 tasks	30.35	38.82	+8.47	0.188
Gemini-3.8-Flash	Diabetes readmission	200	51.00	51.50	+0.50	1.000
Gemini-3.8-Flash	GeneTuring	150	17.82	35.46	+17.64	7.45e-09
Gemini-3.8-Flash	ODIR5K	203	15.27	33.50	+18.23	5.73e-08
Gemini-3.8-Flash	ICU SOFA	200	32.75	34.00	+1.25	0.451
Gemini-3.8-Flash	MIMIC-ICD	201	50.72	48.63	-2.09	0.875
Gemini-3.8-Flash	MedCalc	200	85.00	90.00	+5.00	0.041
Gemini-3.8-Flash	Macro-average (six tasks)	6 tasks	42.09	48.85	+6.75	0.156
Qwen3.5-Plus	Diabetes readmission	200	13.00	42.50	+29.50	7.25e-10
Qwen3.5-Plus	GeneTuring	150	13.87	24.52	+10.64	1.52e-05
Qwen3.5-Plus	ODIR5K	203	23.65	37.44	+13.79	4.26e-06
Qwen3.5-Plus	ICU SOFA	200	35.75	43.75	+8.00	0.006
Qwen3.5-Plus	MIMIC-ICD	201	56.10	48.57	-7.53	1.52e-05
Qwen3.5-Plus	MedCalc	200	84.50	79.00	-5.50	0.099
Qwen3.5-Plus	Macro-average (six tasks)	6 tasks	37.81	45.96	+8.15	0.156
Qwen3.8-27B	Diabetes readmission	200	15.00	48.50	+33.50	1.36e-11
Qwen3.8-27B	GeneTuring	150	12.08	23.43	+11.35	0.000534
Qwen3.8-27B	ODIR5K	203	34.48	38.42	+3.94	0.185
Qwen3.8-27B	ICU SOFA	200	15.75	43.50	+27.75	1.81e-18
Qwen3.8-27B	MIMIC-ICD	201	51.20	50.55	-0.65	1.000
Qwen3.8-27B	MedCalc	200	62.00	77.00	+15.00	0.000535
Qwen3.8-27B	Macro-average (six tasks)	6 tasks	31.75	46.90	+15.15	0.062

Supplementary Table S9. Eight reported donor–target experience transfers

The eight reported donor–target pairs correspond to main Fig. 3c. Frozen source libraries are mounted on target tests using the accumulation backbone. Differences are computed before rounding. Uncorrected two-sided exact McNemar tests compare positive-score outcomes with the target Base.

Donor library	Target task	n	Base %	Library %	Gain pp	p vs Base
GeneTuring	Diabetes readmission	200	24.50	54.00	+29.50	2.73e-09
ICU SOFA	Diabetes readmission	200	24.50	52.00	+27.50	1.72e-08
ICU SOFA	MedCalc	200	68.00	71.50	+3.50	0.419
ODIR5K	MedCalc	200	68.00	71.00	+3.00	0.480
Diabetes readmission	ICU SOFA	200	37.00	39.75	+2.75	0.590
GeneTuring	ICU SOFA	200	37.00	38.00	+1.00	0.798
MedCalc	ICU SOFA	200	37.00	38.00	+1.00	0.389
Diabetes readmission	GeneTuring	150	18.71	18.98	+0.28	1.000

Supplementary Table S10. JevGate training and deployment configuration

JevGate updates the LoRA and pointer-head parameters while keeping the base weights frozen. Agreement evaluation uses a threshold of 0.5; deployment uses fitted task-specific thresholds.

Setting	**Final configuration**
Base model	Qwen3.5-9B-Base; frozen base weights
Initialization	Kev-9B LoRA and pointer-head initialization
Trainable parameters	LoRA (rank 16) and pointer head (dimension 256)
Objective	Binary cross-entropy; equal weight per sampled record
Optimizer	AdamW; weight decay 0.01
Peak learning rate	2×10^{-5}
Schedule / clipping	OneCycle / gradient norm 1.0
Batch	1 per process $\times$ 2 GPU processes $\times$ 8 accumulation = 16 effective
Epochs / selected checkpoint	3 epochs / optimization step 2000
Input character limits	Prompt 9500; reference answer 1000; agent answer 2000; error hint 200
Encoded state budget	3072 tokens
Merged pool	13315
Sampled training records	11290
Admission / rejection labels	7903 / 3387
Evaluation threshold	0.5
Deployment threshold	Task-specific fitted admission thresholds

Supplementary Table S11. Balanced admission-quality comparison before and after fine-tuning

Each task contributes 100 candidates with 50 reference-admission and 50 reference-rejection labels. Agreement uses probability ≥ 0.5 for admission. AUC uses average ranks for ties. The pooled row is computed across all 600 candidates, not by averaging task-level AUCs.

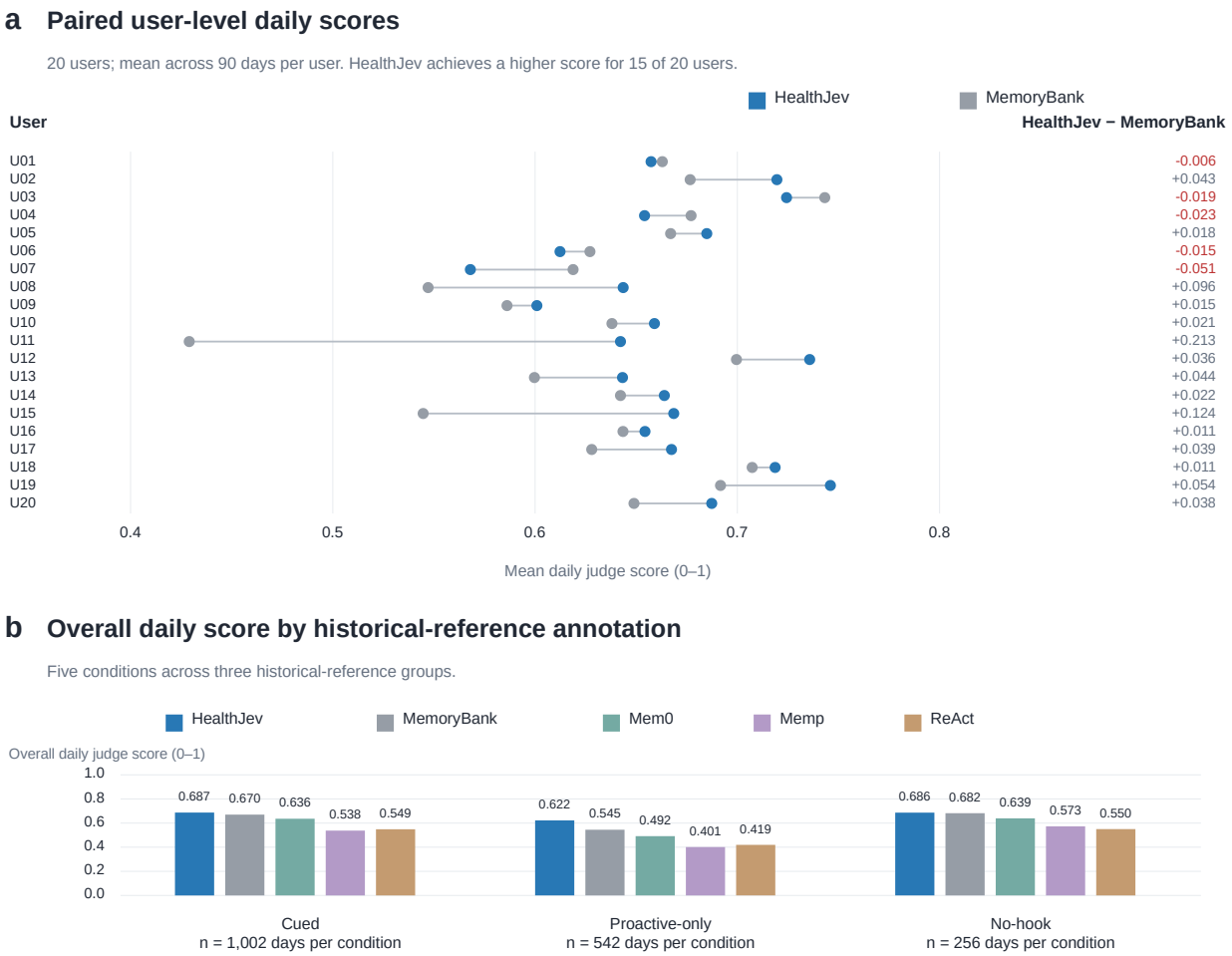
Task	n	Admission / rejection labels	Agreement before	Agreement after	AUC before	AUC after
Diabetes readmission	100	50/50	0.570000	0.710000	0.696800	0.794000
GeneTuring	100	50/50	0.550000	0.760000	0.727600	0.815200
ODIR5K	100	50/50	0.540000	0.710000	0.752200	0.789600
ICU SOFA	100	50/50	0.500000	0.710000	0.601000	0.772400
MIMIC-ICD	100	50/50	0.520000	0.760000	0.876800	0.919200
MedCalc	100	50/50	0.510000	0.760000	0.657600	0.841400
Pooled (600 candidates)	600	300/300	0.531667	0.735000	0.733072	0.809933

Supplementary Table S12. Recorded admission-decision scale

Across 2,983 recorded decisions, 2,440 candidates are admitted (81.8%). The pooled median and 90th-percentile decision latency are 86 and 184 ms after rounding. Mean recorded usage is approximately 1,000 input and 22 output tokens. The denominator includes hard-rule and zero-token decision records, rather than counting serving calls alone. Measurements describe admission decisions, not total agent-response latency.

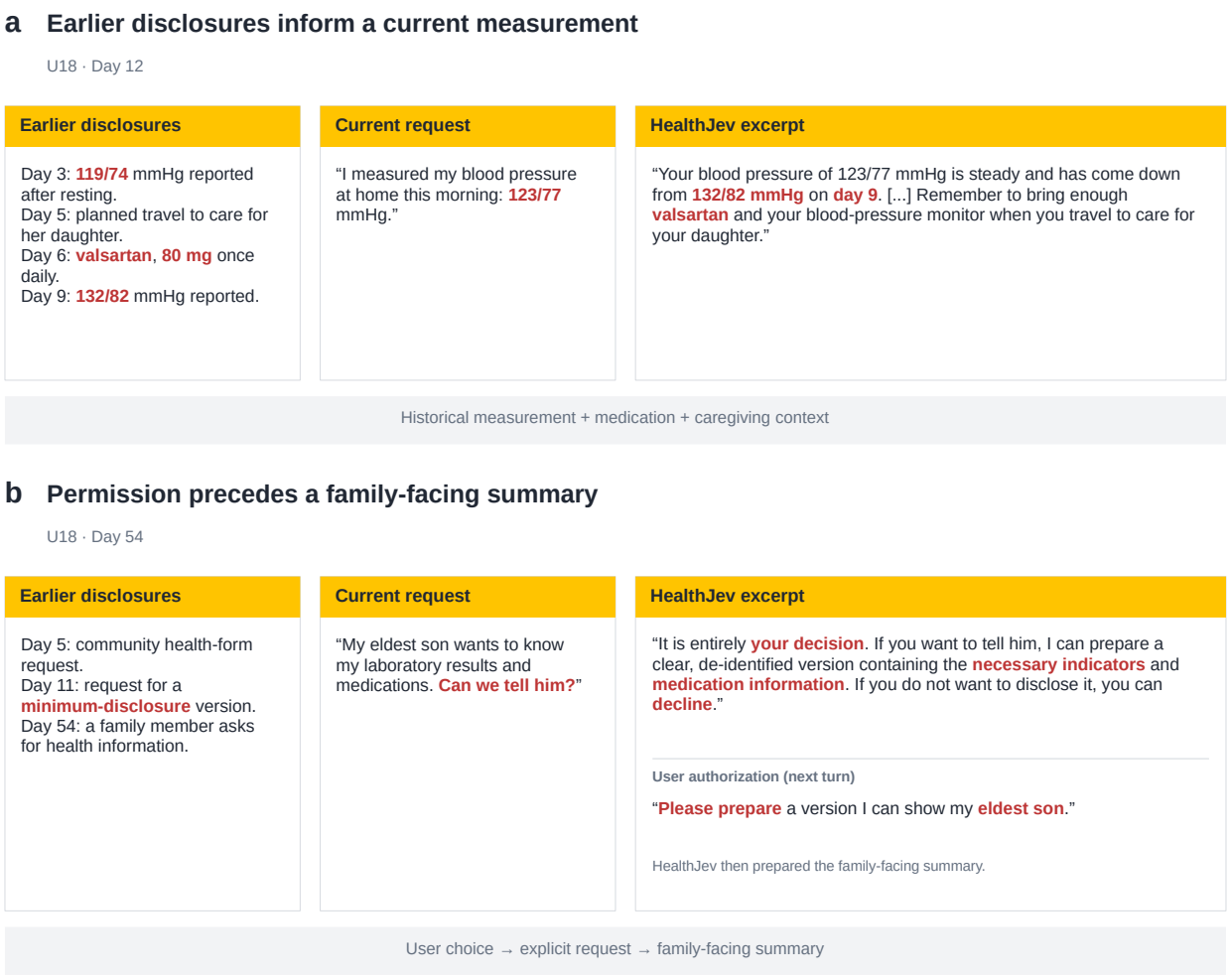
Task	Recorded decisions	Admitted	Rejected	Admission %
Diabetes readmission	581	423	158	72.8
GeneTuring	328	280	48	85.4
ODIR5K	526	410	116	77.9
ICU SOFA	577	458	119	79.4
MIMIC-ICD	549	474	75	86.3
MedCalc	422	395	27	93.6
Pooled	2983	2440	543	81.8

Supplementary Figure S1



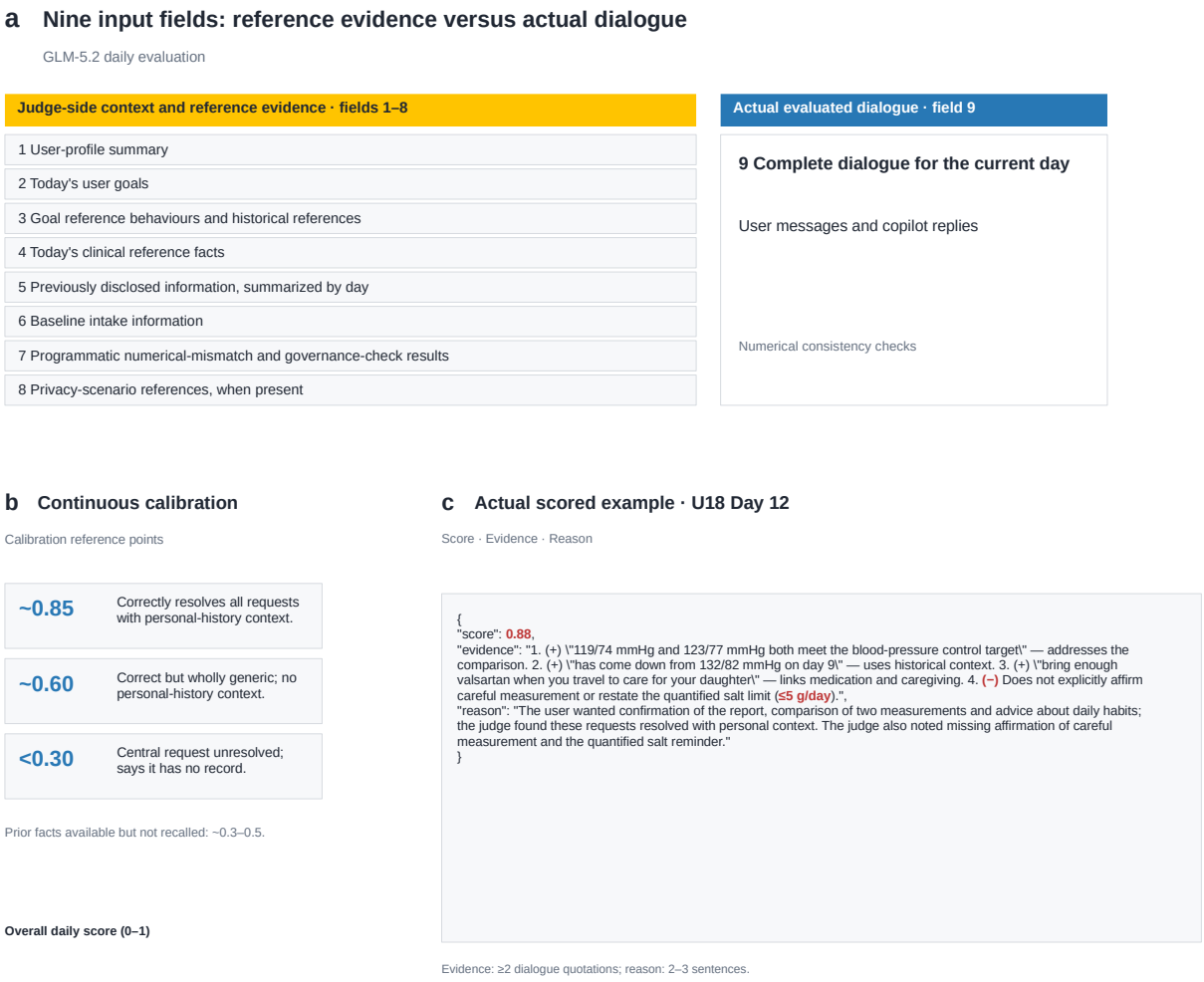
User pairing and descriptive day subgroups. User-level and reference-context performance in the final longitudinal evaluation. a, Mean daily scores for HealthJev and MemoryBank for each of 20 users over 90 days; lines connect matched users. b, Mean overall daily judge score for all five conditions on days with cued references (1,002), proactive-only references (542) or no references (256). Days with any cued reference enter the cued group. Subgroup bars are descriptive overall scores, not separate recall-accuracy estimates.

Supplementary Figure S2



Personal context and disclosure permission. Selected longitudinal dialogue excerpts illustrating continuity and disclosure control. a, Earlier blood-pressure, medication and caregiving disclosures inform the response to a current measurement. b, The user is offered a choice before requesting a family-facing summary. The panels show dialogue excerpts from U18 on days 12 and 54; [...] marks omissions. The final grey annotation records preparation of the family-facing summary.

Supplementary Figure S3



Blinded daily-judge contract. Blinded daily judging protocol and evidence-linked scoring. a, Dialogue and reference evidence supplied to the judge. b, Continuous scoring and illustrative calibration anchors. c, Strict JSON output illustrated by the actual day-12 score of 0.88 for U18. The protocol produces one overall daily score.

Supplementary Figure S4

Reviewer prompt

Prompt | a. Role and all nine inputs

System message
You are a quality evaluator for a personal health concierge service. Output only **strict JSON**.

Role
You are a quality evaluator for a personal health concierge service. Your task is to read the entire conversation between the user and the AI health concierge for the day, then assign a **continuous score** between 0 and 1 measuring the extent to which the AI concierge met the user's needs that day. You evaluate **quality only**; you do not know, and do not need to know, which system is being evaluated.

Inputs
1. User profile summary: {profile_summary}
2. User goals for the day: {goals_content}
3. Ground truth for evaluating the goals: {expected_behavior_and_cross_refs}
4. Ground-truth facts for the day (hospital documents / fact history, including dual units): {ledger_truth}
5. Summary of historical disclosures (by day): {disclosure_summary}
6. User intake archive (Day 0 baseline laboratory results / current medications / medical history): {day0_archive}
7. Programmatic validation results: {value_mismatch_list}
8. Ground truth for privacy events (if any): {privacy_truth}
9. Complete conversation for the day: {transcript}

Prompt | b. Judging criteria and output

Core scoring question
Go through what the user said that day, item by item: as a personal health concierge, did the response, overall, meet the user's needs?
The user reports a situation: did they receive an interpretation and advice tailored to them personally (taking account of medical history, medications, previous measurements, and life context), rather than generic health information that would apply equally to any user?
The user asks about past events, data, or agreements: did the concierge respond accurately? The user comes to the concierge precisely because it should remember their circumstances. If this information **actually exists** in historical disclosures or the archive, but the concierge merely says "I do not know / please provide it," such a response is useless to the user and should receive a low score.
The concierge proactively references correct details of the user's personal history, gives advice suited to their condition, flags risks, and communicates naturally: high score. Responses that fail to address the question, offer formulaic dismissals, repeatedly request information the user has already provided, or reveal internal operational processes: low score.
Content should be safe and reliable: **medical errors**, **unsafe advice**, or privacy problems (such as disclosing the user's health information to a third party **without authorization**) likewise constitute serious failures to meet the user's needs.

Calibration references
A conversation that correctly resolves all of the user's requests for the day and provides personalized responses informed by personal history / the archive should typically score around **0.85**. A conversation that is correct but makes no use of personal history at all (generic responses applicable to any user) should typically score around **0.6**. A conversation that fails to resolve the core request and merely says "there is no record" should typically score **below 0.3**. Your scores should remain broadly consistent with the scale indicated by these reference points.

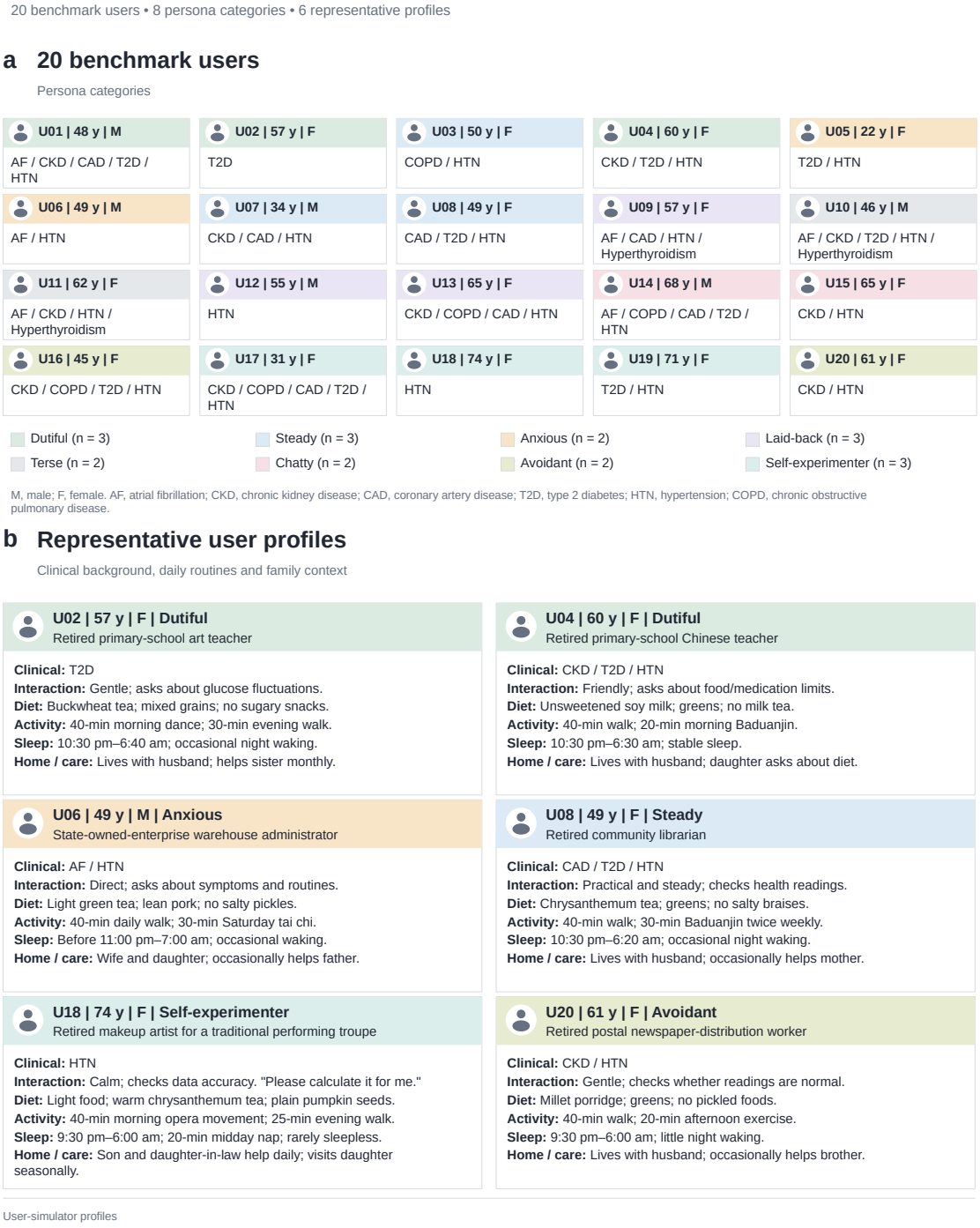
An important comparison principle
A response informed by the user's actual history (even if a few details are imperfect) > a generic but correct response > a response that knows nothing about information the user has already provided and merely says "I do not know." The user wants "a concierge who remembers me," not "a polite concierge who knows nothing about me."

Output (strict JSON)
{{"score": **continuous score** from 0 to 1, "evidence": "Quote key excerpts **verbatim** from the conversation (**at least 2**), indicating why each supports the score or warrants a deduction", "reason": "**2-3 sentences**: what the user wanted that day, whether the concierge provided it, and what was good / poor"}}

Discipline
Assign the **continuous score** directly; **do not use fixed scoring bands**. Score the actual quality of the conversation. Positive performance supported by concrete evidence (requests correctly met and personal history incorporated) should receive the corresponding score. Evidence must quote the original conversation **verbatim**. You do not know the provenance of the system being evaluated. The score must be **consistent with the evidence and reason**; otherwise, the output is invalid.

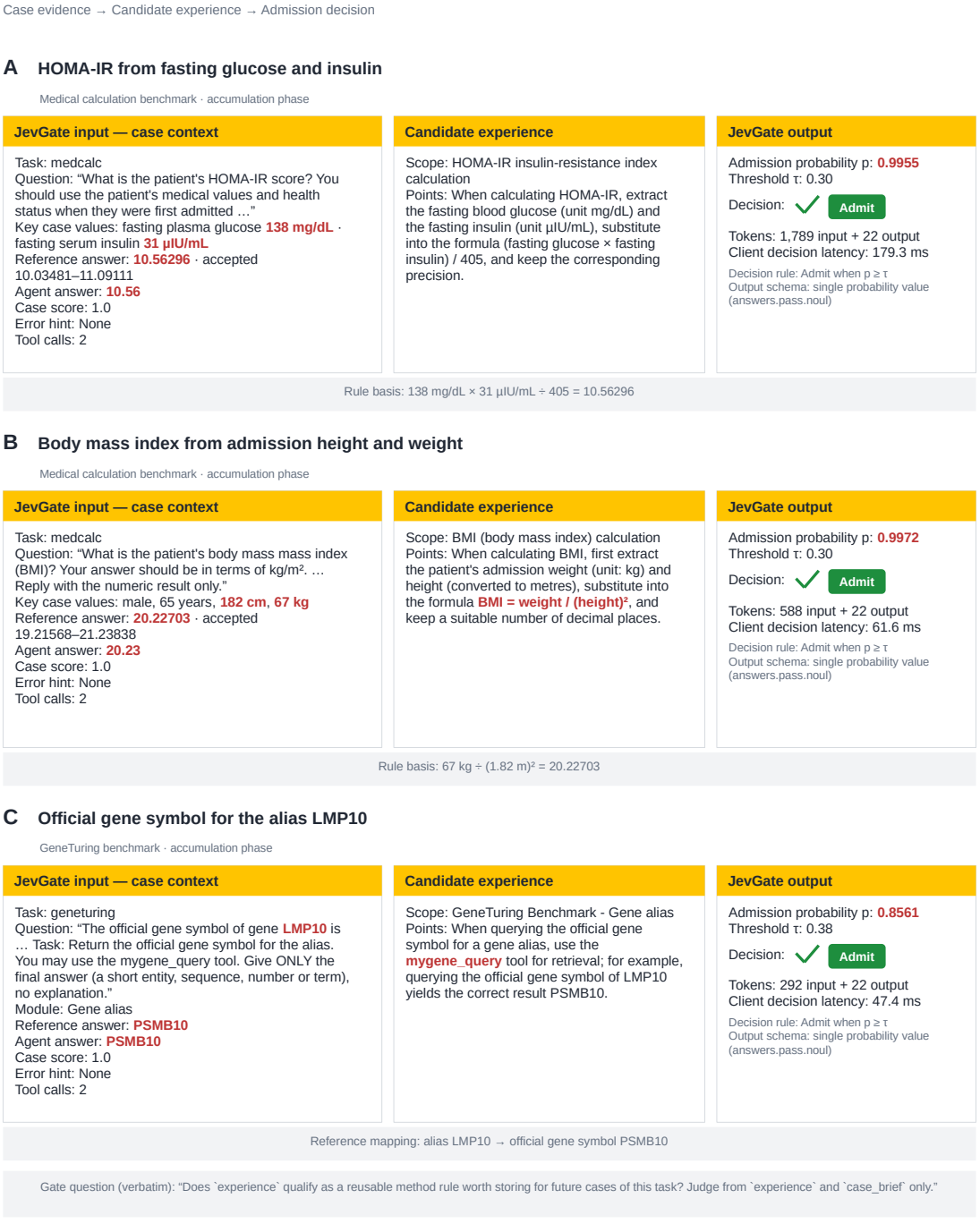
Blinded daily judging prompt. Excerpts from the blinded daily-judging prompt. a, System message, evaluator role and nine input fields. b, Scoring criteria, calibration references, comparison principle, JSON output contract and evidence discipline. The prompt retains all nine input placeholders.

Supplementary Figure S5



Longitudinal benchmark user-simulator profiles. a, Overview of the 20 benchmark users, ages 22–74, with their health backgrounds and eight persona categories. b, Six representative profiles detail occupation, interaction style, diet, activity, sleep, household and caregiving context. Colours identify the persona categories used by the user simulator.

Supplementary Figure S6



JevGate admission examples. JevGate admission decisions for reusable calculation and gene-reference rules. a, HOMA-IR unit and formula handling. b, BMI height conversion and calculation. c, Gene-alias retrieval of the official symbol PSMB10. Each row shows case-context inputs, the candidate experience and its recorded admission probability p, task threshold and decision. Case score is the programmatic score of the originating answer; p is the probability of admitting the candidate experience. Inputs and experiences are presented as excerpts. Token usage and client-side decision latency are recorded for the corresponding original full gate call.