Frontier LLMs require explicit clinical context to avoid training-data anchoring for insulin pump settings: a pre-registered exploratory study

Timothy Street

April 2026

Running head: LLMs require context to avoid pump-settings anchoring

Timothy Street

Independent researcher, UK — April 2026 preprint

Pre-registration: OSF Registries, DOI 10.17605/OSF.IO/Q4UX9

Abstract

Objective. To test whether frontier large language models derive insulin pump settings from patient data or from training-data priors, and whether providing explicit clinical context overcomes any observed anchoring. Exploratory study, N=3 AID users.

Methods. Five models (Claude Sonnet 4.6, Claude Opus 4.7, GPT-5.4, Gemini 2.5 Pro, Gemini 3.1 Pro Preview — pre-release, 34-74% output failure rate) were each prompted 50 times with seven days of 5-minute CGM and hourly pump data from three users with ISF 45-163 mg/dL/IU and ICR 7.5-26 g/IU (1,350 calls). AID system type and current settings were deliberately withheld to isolate data-based reasoning. Six hypotheses were pre-registered before analysis. Post-hoc analyses tested training-prior anchoring; a context ablation (600 additional calls) tested whether providing the user's AID system and current settings reduced anchoring; citation accuracy was verified by exact-SGV match against raw 5-minute data.

Results. Of six pre-registered hypotheses, one was fully confirmed (DIA collapse to 4-5 hours, 100% of iterations), two partially confirmed, two untestable, and one not confirmed. Without clinical context, models' recommendations were closer to their own cross-user average than to the user's actual profile in 90% of comparisons. Providing AID system context and current settings substantially reduced this anchoring: DIA was corrected completely (5h → 9h), ICR improved from 0-22% to 58-98% within ±10%, and ISF improved in most cells. However, the improvement appeared to reflect anchor substitution rather than data-driven reasoning — the model explicitly stated it was echoing the provided settings because it could not derive ICR from the data. Basal — the most data-driven setting — sometimes worsened with context. Citation accuracy was unaffected by context: 30-70% of cited glucose values were misattributed from elsewhere in the dataset regardless of whether the model knew the user's settings.

Conclusions. In this exploratory three-user study, frontier LLMs required explicit clinical context (AID system type and current settings) to produce pump recommendations approaching the user's validated profile. Without this context, models anchored on training-data priors. With context, recommendations improved substantially but through anchor substitution rather than clinical reasoning from the data. Settings anchoring and citation misattribution were independent failure modes. These findings require confirmation in a larger, more diverse sample.

---

1. Introduction

Large language models are increasingly marketed as clinical reasoning tools, and AI-powered "diabetes advisors" — from Custom GPTs to commercial prototypes — already let users paste device data and receive insulin-dosing recommendations. For an adult with Type 1 diabetes on an automated insulin delivery (AID) system, the settings in question — basal rate, insulin sensitivity factor (ISF), and insulin-to-carbohydrate ratio (ICR) — directly control insulin delivery.

The implicit promise of these tools is that the model *reads* the patient's data and *derives* personalised settings from it. This paper tests that promise.

We began with a different set of expectations. Based on a pilot study on a single user, we pre-registered six hypotheses focused on per-model reproducibility rankings, citation hallucination rates, and temperature sensitivity — treating each model as a distinct clinical tool to be characterised independently (OSF DOI: 10.17605/OSF.IO/Q4UX9). When we expanded to three users, a different picture emerged. The per-model rankings shifted between users in ways the pre-registration did not anticipate. More importantly, the *cross-user pattern* — all models converging on similar values regardless of whose data they were given — pointed to a single underlying mechanism: the models' outputs appeared to be shaped more by their training data than by the patient data in the prompt.

This training-prior anchoring hypothesis was not pre-registered. It emerged from the data and was subsequently tested with three independent analyses and a context ablation study. The paper reports both the pre-registered results and the exploratory findings, clearly separated.

2. Methods

2.1 Users and data

Three adults with Type 1 diabetes on AID systems contributed seven days (April 9-15 2026) of anonymised CGM and pump telemetry via their Nightscout instances at native 5-minute CGM resolution with 168 hourly device summaries. User 2's duplicate CGM entries (from parallel Dexcom Share and Trio uploads, identified by near-identical timestamps 1-2 seconds apart with different `_id` formats) were de-duplicated by retaining one reading per 5-minute bucket (`bucket = epoch_ms // 300000`), reducing 3,076 raw entries to 1,992.

	User 1	**User 2**	**User 3**
AID system	Boost fork of AAPS	Trio	Trio
CGM readings	1,969 (97.7%)	1,992 (98.8%)	2,019 (100.1%)
Device hours	168/168	168/168	168/168
TDB (U/day)	8.11	15.20	18.65
DIA (hours)	10	9	9
Basal (U/hr)	0.275-0.396	0.550-0.750	0.700-0.850
ISF (mg/dL/IU)	91-163	90-115	45 (flat)
ICR (g/IU)	21-26	15 (flat)	7.5 (flat)
Carb entries	0	0	25

Users 2 and 3 have similar TDB but markedly different ISF and ICR, enabling a test of whether models respond to insulin-sensitivity characteristics independently of basal volume. All profiles were iteratively tuned by the users themselves based on observed glucose outcomes — standard practice in T1D pump management, particularly among AID users with continuous feedback from CGM and algorithm behaviour. All three users had stable HbA1c and time-in-range on these settings. No meal, bolus, or system-identification data were included in the primary analysis.

2.2 Prompt design

A structured prompt defined the advisor role ("Balanced Clinical Consultant"), required a JSON schema with 24 hourly entries for basal/ISF/ICR plus per-hour confidence, citations with timestamps, mechanical events, unknown variables, and a self-check step. Anti-hallucination instructions forbade inventing data not present in the input. The prompt did **not** identify the AID system, insulin type, or user's current settings. This was a deliberate design choice to establish a baseline: can models reason from data alone? This baseline is necessary to make the context ablation (§5) interpretable —

without it, we cannot determine whether context provision improves outcomes through genuine reasoning or through anchor substitution. However, this baseline does not represent realistic deployment conditions, where users would typically provide their system and settings. The ablation study (§5) tests the realistic scenario and should be considered alongside the primary analysis when evaluating clinical implications.

2.3 Models

Model	Provider	Temperature	Cost (est.)
Claude Sonnet 4.6	Anthropic	0.00, 0.01	~$27
Claude Opus 4.7	Anthropic	not supported	~$123
GPT-5.4	OpenAI	0.00, 0.01	~$42
Gemini 2.5 Pro	Google	0.00, 0.01	~$10
Gemini 3.1 Pro Preview*	Google	0.00, 0.01	~$10

*Pre-release; behaviour may change before GA.

50 iterations per temperature per user per model (Opus: 50 per user, no temperature). Total: 1,350 calls.

2.4 Pre-registration

Six directional hypotheses were derived from a pilot study and pre-registered on OSF before the present analysis. H1: GPT-5.4 lowest basal CV, Gemini 2.5 highest (≥2/3 users). H2: Gemini 2.5 basal recommendations >20% above profile in >20% of values (≥2/3 users). H3: all models DIA ≤6h in ≥90% of iterations (all cells). H4: Sonnet <1% out-of-window citations all users, Gemini 3.1 >10% all users. H5: Gemini models detect <20% of profile changes. H6: Gemini 3.1 ISF CV ratio ≥3× between temperatures (≥2/3 users).

2.5 Metrics

Per-hour CV (reproducibility), closeness to profile (% within ±10%, signed bias), citation veracity (timestamp validity and exact-SGV match against raw 5-minute data), output validity (% parseable JSON with 24 hourly entries).

2.6 Post-hoc analyses (exploratory)

The following analyses were developed after the pre-registered hypotheses were tested. They are clearly separated from the confirmatory results and should be treated as exploratory:

(a) Training-prior anchoring tests. Three independent tests: adjustment magnitude across users, distance-to-model-average vs distance-to-profile, and null-prompt elicitation of “typical” settings. Analysis script: `anchoring_analysis.py`.

(b) Context ablation. The prompt was modified to include the user's AID system type and current settings (DIA, ISF, ICR, basal range). 600 additional queries across Sonnet, GPT-5.4, Gemini 2.5 Pro, and Gemini 3.1 Pro Preview (50 per user per model at t=0.0).

(c) Citation classification. Each of 10,514 citations classified as correct (exact SGV match), misattributed (real value at wrong timestamp), invented (value not in dataset), or fabricated timestamp. Analysis script: `citation_classification.py`.

(d) Semantic similarity. Summary-field embeddings (all-MiniLM-L6-v2) with within-cell cosine similarity and UMAP clustering. Analysis script: `semantic_similarity.py`.

(e) Treatment data test. Bolus and carb data added for User 3 (the only user with carb entries). 50 iterations, Sonnet, t=0.0.

3. Results: Pre-registered hypotheses

3.1 H3 confirmed: DIA collapse is universal

All five models recommended DIA ≤6 hours in 100% of parseable iterations, for all three users, at both temperatures (Figure 5). Median DIA was 4.0-5.0 hours against user profiles of 9-10 hours. This is the only pre-registered hypothesis fully confirmed.

3.2 H1 partially confirmed: reproducibility rankings

Model	User 1 CV	User 2 CV	User 3 CV	Mean
Claude Sonnet 4.6	**2.7%**	**3.0%**	**3.9%**	**3.2%**
Claude Opus 4.7	2.8%	6.3%	8.9%	6.0%
GPT-5.4	6.5%	3.4%	9.3%	6.4%
Gemini 2.5 Pro	10.0%	9.2%	13.9%	11.0%
Gemini 3.1 Pro Preview	7.3%	0.0%*	0.0%*	—

*Artefact of low valid-response count producing near-identical truncated outputs.

H1 predicted GPT-5.4 lowest and Gemini 2.5 highest in ≥2/3 users. Gemini 2.5 was consistently highest (confirmed), but **Sonnet was lowest, not GPT-5.4** (GPT-5.4 ranged 3.4-9.3% across users). H1 is partially confirmed: the least-reproducible model is stable, but the most-reproducible model was not correctly predicted.

3.3 H2 not confirmed: recommendations significantly higher than profile

H2 predicted Gemini 2.5 Pro basal recommendations >20% above the user's profile in >20% of values, in ≥2/3 users. Results: User 1: 78% (passes), User 2: 18% (fails the >20% threshold), User 3: 0%. **Only 1/3 users exceeded the pre-registered threshold.** The effect is user-dependent, not a universal Gemini 2.5 property — it occurs when the user's settings are below the model's training-data centre of gravity but not when they are above it.

3.4 H4 partially confirmed: citation veracity

Sonnet: 0.0%, 0.1%, 0.0% out-of-window — confirmed <1% across all users. Gemini 3.1 Pro Preview: 4.4%, 2.0%, 54.6% — exceeded 10% only for User 3 (where truncation failures were most severe). **H4 is partially confirmed: the Sonnet prediction held; the Gemini 3.1 prediction did not.**

3.5 H5 inconclusive, H6 not testable

No profile changes occurred in any user's data window (H5 precondition not met). Gemini 3.1's escalating truncation failures (34-74% at t=0.01) prevented reliable CV computation at both temperatures (H6 not testable).

3.6 Summary

H#	Prediction	Result
H1	Reproducibility rank order	Partially confirmed
H2	Gemini 2.5 recommendations significantly higher	**Not confirmed** (1/3 users)
H3	DIA ≤6h all cells	**Confirmed** (100%)
H4	Citation veracity divergence	Partially confirmed
H5	Gemini misses profile changes	Inconclusive
H6	Gemini 3.1 temperature sensitivity	Not testable

One of six hypotheses fully confirmed. The per-model framing — treating each model as a distinct tool to be ranked — was incomplete. The next section examines the cross-model pattern that the pre-registration did not anticipate.

4. Results: Training-prior anchoring (exploratory)

4.1 Cross-model convergence

All five models recommended ISF in the range 50-100 mg/dL/IU and ICR in the range 10-16 g/IU regardless of user, despite the reference ISF spanning 45-124 mg/dL/IU and ICR spanning 7.5-26 g/IU. User 3 (ISF 45, ICR 7.5) scored 0% closeness-to-profile (within ±10%) for every model except Sonnet (18.6%). Users 2 and 3 have similar TDB (15.2 vs 18.7 U/day) but very different outcomes, confirming that ISF/ICR distance from the models' training-data centre of gravity — not basal volume — is the primary predictor of failure.

4.2 Adjustment test

The three users' reference ISF spans 79 mg/dL/IU. The models' recommended ISF spans:

Model	ISF spread	% of reference	ICR spread	% of reference
Sonnet 4.6	63	80%	5.9	35%
Opus 4.7	8	**10%**	0.6	**4%**
GPT-5.4	21	26%	4.1	24%
Gemini 2.5	20	26%	1.7	10%
Gemini 3.1	2	**3%**	1.4	8%

Only Sonnet adjusts ISF by more than a third of the reference range. Every model adjusts ICR by less than a third (Figure 6). Opus 4.7, at 4.6× the cost of Sonnet, adjusted ISF by only 8 mg/dL/IU and ICR by 0.6 g/IU — recommending near-identical settings for all three users.

4.3 Distance test

Across 30 model × user × setting (ISF + ICR) comparisons, the recommendation was closer to the model's own cross-user average than to the user's actual profile in **27/30 cases (90%)**.

4.4 Null-prompt test

With no patient data, Sonnet and GPT-5.4 reported "typical" ISF midpoints of 50-60 mg/dL/IU, ICR midpoints of ~12.5 g/IU, and DIA of 4 hours. These closely match the values they recommend when given data.

4.5 Opus 4.7: larger model, stronger anchoring

Opus 4.7 costs 4.6× more than Sonnet 4.6 and showed stronger anchoring on every axis: lower ISF adjustment (10% vs 80%), lower closeness-to-profile (32.2% vs 41.7%), larger bias (−28% vs −8.4%), and higher citation fabrication (7.32% vs 0.03%). A more capable model from the same vendor appears to anchor more firmly, not less.

5. Results: Context ablation (exploratory)

The primary analysis (§3-4) established that models anchor on training priors when clinical context is withheld. However, this is not how these tools would be used in practice — real-world users would typically provide their AID system and current settings. This section tests the realistic scenario: does providing context overcome the anchoring, and if so, through what mechanism?

5.1 Method

The prompt was modified to include the user's AID system type (Trio/AAPS, oref-based), current DIA, ISF, ICR, and basal range, with the instruction: "Use these current settings as a reference point." 600 queries across Sonnet, GPT-5.4, Gemini 2.5 Pro, and Gemini 3.1 Pro Preview (50 per user per model at t=0.0).

5.2 DIA — pure training-data recall

Every model × user combination snapped from 3.5-6h to exactly 9-10h across all four models (Figure 7). The training prior was completely overridden by explicit context, confirming DIA collapse was entirely attributable to missing system context.

5.3 ICR — anchor substitution, not reasoning

	User 1 (no carbs)	**User 2 (no carbs)**	**User 3 (25 carb entries)**
Sonnet: without → with	3.8% → **58.1%**	0.3% → **84.2%**	0.5% → **96.4%**
GPT-5.4: without → with	2.8% → **58.2%**	22.2% → **98.0%**	0.0% → **70.7%**
Gemini 2.5: without → with	2.5% → **31.1%**	68.9% → **32.0%** ⚠	0.0% → **83.3%**
Gemini 3.1: without → with	0.0% → **54.3%**	no valid responses	0.0% → **0.0%** ⚠

ICR improved dramatically for most model × user cells despite Users 1 and 2 having **no carb entries or meal boluses** in the data. Two anomalies: Gemini 2.5 User 2 worsened (68.9% → 32.0%) and Gemini 3.1 User 3 was unchanged at 0% — context did not help these cells. The model cannot

validate ICR against actual meals; it echoes the stated value. Sonnet confirmed this explicitly: *"ICR values are kept close to current settings (20-25 g/IU) as no meal records are available"* (User 1 with context) and *"Without bolus data, ICR cannot be meaningfully calculated from this dataset"* (User 1 without context). The improvement reflects anchor substitution — replacing the training-data anchor with the stated-settings anchor — not data-driven reasoning. The Gemini 2.5 anomaly on User 2 (ICR worsened from 68.9% to 32.0%) demonstrates that context does not guarantee improvement.

5.4 ISF — mixed signal

ISF improved in most cells (+15 to +52 percentage points) but GPT-5.4 User 2 worsened (41.2% → 27.7%) and Gemini 3.1 User 3 collapsed (50.0% → 0.0%). ISF sits between DIA (pure anchor) and basal (mostly data-driven): the model partially reasons from correction-response patterns in the CGM data, so context helps but doesn't fully determine the output.

5.5 Basal — the most data-driven setting

Sonnet User 1 worsened (55.7% → 46.4%). Gemini 2.5 User 1 unchanged (11.9% → 11.0%). Gemini 3.1 User 1 worsened (61.3% → 56.0%). Other cells improved modestly. Basal is the setting most directly derivable from CGM patterns (overnight fasting, temp-basal delivery). Providing context can override data-based reasoning, occasionally making results worse.

5.6 The anchoring spectrum

These results reveal a spectrum of data-dependence:

- **DIA**: pure anchor. Zero data signal. Context completely overrides.
- **ICR**: mostly anchor. Minimal data signal (no meals for 2/3 users). Context substitutes one anchor for another. Model admits it cannot derive ICR from the data.
- **ISF**: mixed. Some data signal (correction responses). Context helps AND the model partly reasons from data.
- **Basal**: most data-driven. Strong CGM signal. Context sometimes worsens by overriding data-based reasoning.

This spectrum is visualised in Figure 8.

5.7 Gemini 3.1 Pro Preview: context does not fix structural failures

Gemini 3.1 Pro Preview (pre-release) continued to exhibit output-truncation failures even with full context: 48/50 valid for User 1, 0/50 for User 2, and 23/50 for User 3. Where it did produce valid output, context improved DIA (5h → 9-10h) and ICR for User 1 (0% → 54.3%), but ISF for User 3 collapsed from 50% to 0% with context — a worsening. This demonstrates

that context provision cannot overcome architectural output-reliability issues, and that the benefits observed for the GA models do not generalise to this pre-release model.

5.7 Citations unaffected by context

Sonnet's exact-SGV match rate changed from 38.1% to 41.6% for User 3 (+3.5%). Citation misattribution is architecturally distinct from settings anchoring. Additionally, 65% of User 1's citations referenced carbs, meals, or postprandial patterns despite zero carb entries in the data — the model generates meal-related reasoning from its training distribution, not from the data.

6. Results: Citation verification

6.1 Exact-SGV match

Models were given exact integer glucose values. Citation accuracy (pooled across all users):

Model	Correct	Misattributed	Invented	Fab. timestamp	n
GPT-5.4	58.3%	41.4%	0.3%	0.0%	2,915
Sonnet 4.6	51.6%	48.3%	0.1%	0.0%	3,423
Opus 4.7	30.8%	61.4%	0.5%	7.3%	2,488
Gemini 2.5	28.4%	68.1%	1.4%	2.1%	1,188
Gemini 3.1	65.8%	18.2%	0.4%	15.6%	500

No model correctly cites more than two-thirds of glucose values (Figure 9).

6.2 Error classification

Among models with valid timestamps, 98-100% of incorrect values are misattributions — the cited value exists elsewhere in the 7-day dataset but is attributed to the wrong timestamp. For Opus 4.7 and Gemini 3.1 (which also fabricate timestamps), misattributions account for 89% and 54% of errors respectively. Pure invention of glucose values is rare (0.1-1.4%). The dominant error pattern — cross-referencing a real value from the wrong row — is consistent with probabilistic attention over the context window rather than precise row-level indexing.

7. Results: Additional analyses

7.1 Treatment data

Adding bolus and carb data for User 3: ISF improved partially (+23% → +17% bias), ICR worsened (+28% → +34% bias), basal and DIA unchanged. More data does not systematically improve recommendations.

7.2 Gemini 3.1 Pro Preview truncation

Output validity degraded from 50/50 (User 1, t=0.0) to 13/50 (User 3, t=0.01) — a 74% failure rate. The truncation is data-dependent and temperature-dependent. This is a pre-release model; findings may not persist to GA.

7.3 Semantic similarity

Narrative reproducibility (within-cell cosine similarity at t=0.0) was uncorrelated with numeric reproducibility (Spearman rho = −0.02, p=0.95). Gemini 2.5 produced the most narratively stable summaries (cosine 0.90) despite the highest numeric CV (11%). UMAP clustering separated responses by source model with 89% accuracy, indicating each model has a distinct rhetorical fingerprint.

8. Discussion

8.1 From model comparison to mechanism

We entered this study expecting to characterise five models as distinct clinical tools. The pre-registered hypotheses reflected this framing. What the data showed was that the differences *between models* were less important than the similarity in how *all models* behaved: they converged on the same narrow range of ISF and ICR values regardless of the user data, adjusted minimally between users, and defaulted to textbook DIA.

8.2 Training-prior anchoring

This convergence points to a single mechanism: the models' outputs appear to be shaped more by their training data than by the patient data in the prompt. The evidence is threefold: (a) models adjust ISF by only 3-80% of the reference range between users, (b) 90% of recommendations are closer to the model's own cross-user average than to the user's profile, and (c) the models' stated "typical" settings with no data match what they recommend with data.

The context ablation confirms the mechanism is real and reducible: providing the user's AID system and current settings dramatically improved DIA and ICR. But the improvement mechanism matters. For ICR, the model

explicitly stated it was echoing the provided settings because it could not derive ICR from the data. For basal — the most data-driven setting — context sometimes worsened results by overriding data-based reasoning.

8.3 How LLMs process clinical data

Three architectural properties are relevant:

(a) The context window is not a database. The ~34,000 tokens of CGM data are attended to probabilistically, not queried like rows in a table. This explains the citation misattribution pattern: the model knows approximately what glucose values appear in the data but cannot reliably index a specific value at a specific timestamp.

(b) Training creates probability distributions over outputs. The model has learned that pump settings "typically" look like ISF 50-100 and ICR 10-20. Patient data can shift these distributions, but only if the data signal is strong enough to overcome the prior. For DIA (weak data signal, strong prior) the prior dominates completely. For basal (strong data signal) the model adjusts more but still partially.

(c) The model cannot distinguish "I have evidence" from "I should generate something plausible." It produces ICR recommendations with confidence scores even when it states ICR cannot be derived from the data. It does not refuse — it fills the gap with the strongest available anchor. This is inherent to autoregressive text generation.

8.4 Two independent failure modes

Settings anchoring and citation misattribution are architecturally distinct. Providing context fixes settings anchoring (DIA, ICR) but does not fix citation accuracy (+3.5%). A deployment would need two different safeguards: context provision for settings, and programmatic citation verification for reasoning.

8.5 Implications for deployment

Based on these exploratory findings, these models appear to function as language-generation systems that produce clinical-looking output, rather than clinical reasoning systems that happen to use language. The quality of that output appears to depend on the overlap between the patient's actual settings and the model's training distribution, not solely on the model's analysis of the patient's data — though a larger sample would be needed to confirm this interpretation.

Providing the user's current settings substantially improves closeness-to-profile but may create a confirmation risk: the model echoes the stated settings rather than independently validating them against the data. Whether this matters clinically depends on how such tools are deployed — as a second opinion (where echoing is unhelpful) or as a data-conditioned refinement (where the starting point is the user's existing settings and small adjustments may be valuable).

9. Limitations

Sample size and generalisability. Three users across two AID systems is an exploratory sample. The study cannot disentangle user-identity from AID-system effects, and the threefold ISF/ICR range — while deliberately sought — reflects convenience sampling from the author's network rather than a powered design. No formal power calculation was performed; N=3 was resource-limited. Inferential statistics are not reported because they would be uninformative at this sample size.

Prompt design confound. The primary analysis (§3-4) deliberately omitted the user's current settings, AID system type, and insulin type. The ablation study (§5) demonstrates that providing this context substantially reduces anchoring for DIA and ICR, indicating the primary analysis overstates the severity of anchoring in realistic deployment scenarios where users would provide their AID system and current settings. The no-context condition should be interpreted as a baseline that enables the ablation to distinguish anchor substitution from genuine reasoning, not as a representation of real-world use.

Reference profile validity. User-tuned settings were used as the per-patient reference. Self-titration of pump settings is standard practice in Type 1 diabetes management — many people with T1D iteratively adjust their own basal rates, ISF, and ICR based on observed glucose outcomes, and this is particularly common among AID users who have continuous feedback from their CGM and algorithm behaviour. The three users in this study have stable HbA1c and time-in-range on their current settings, making them reasonable per-patient references. That said, user-tuned settings are not necessarily optimal — if a model recommends different values that would produce better clinical outcomes, it would score poorly on closeness-to-profile despite being clinically appropriate. Future work could compare model recommendations against both user settings and outcomes-based measures (e.g., projected time-in-range from in-silico simulation).

Citation classification. Citation classification was performed by a single analyst using automated scripts (`citation_classification.py`) without inter-rater reliability testing. The scripts are deposited for independent replication.

Model-specific limitations. Opus 4.7 does not support temperature and was tested without it, limiting direct comparison of temperature sensitivity. Gemini 3.1 Pro Preview is pre-release with 34-74% output-truncation failure rates; its results should be interpreted as reflecting pre-release behaviour and may not persist to GA. The interpretation that Opus shows "stronger anchoring" (§4.5) is based on a single model pair and could reflect model-specific properties rather than a general size effect.

Other. The training-prior anchoring analysis (§4) was not pre-registered and should be treated as exploratory. The null-prompt test was conducted on two of five models (Sonnet and GPT-5.4). No per-hour analysis of ISF/ICR was performed. The study does not quantify the ISF/ICR range within which anchoring produces acceptable recommendations. The three users were not selected to maximise diversity but happened to span a threefold ISF range.

10. Future work

A larger study (15-30 users stratified by ISF, ICR, and AID system) would map the boundaries of the safe zone and determine whether the anchoring spectrum (DIA → ICR → ISF → basal) holds across the full range of insulin sensitivity profiles. A designed comparison of prompts with and without system context — rather than our post-hoc ablation — would provide cleaner evidence. Pre-registration of a regression model (closeness ~ ISF + ICR + AID system + model) would be essential.

11. Conclusions

In this exploratory three-user study, frontier LLMs anchored on training-data priors when generating insulin-pump recommendations without clinical context. Providing the user's AID system and current settings substantially reduced this anchoring — correcting DIA completely and improving ICR and ISF — but the model's own reasoning indicated this improvement reflected anchor substitution rather than clinical derivation from the data. Citation misattribution (30-70% of cited glucose values) was unaffected by context provision, representing an independent failure mode. These findings were consistent across five models from three vendors but require confirmation in a larger, more diverse sample. Based on this exploratory evidence, deployment of LLM-based insulin pump advisors should include: (1) explicit provision of AID system context and current settings, (2) post-hoc bounds checking against validated ranges, (3) programmatic citation verification, and (4) validation in larger samples before clinical use.

Data availability

All analysis scripts, 1,350 raw model responses, 600 context-ablation responses, reference profiles, data blocks, and the pre-registration are available at [OSF project URL TBD]. Pre-registration DOI: 10.17605/OSF.IO/Q4UX9.

Funding

None. Estimated API costs: Anthropic ~$150, OpenAI ~$42, Google ~$20. Total ~$212 (primary) + ~$50 (ablation).

Conflicts of interest

The author is an independent researcher and the owner of the Diabettech blog. No funding or in-kind support from the model vendors was received. Claude Code (Anthropic) was used for data analysis, figure generation, and manuscript drafting assistance. All analysis scripts are deposited and can be independently executed using any Python environment; the use of Anthropic's tool did not influence the evaluation methodology or model

selection. Two of the five tested models are Anthropic products; Sonnet 4.6 showed the most adaptation to user data on some metrics, though no model performed adequately for clinical use without context provision and bounds checking. All content was reviewed and verified by the author.

Consent

All three users provided informed consent via written communication for the use of their anonymised device data.

References

1. Street T. *Comparing reproducibility and accuracy of frontier LLMs for carbohydrate estimation from food images.* SSRN preprint (2026).
2. Sng G L Y, Tung J Y M, Lim D Y Z, Bee Y M. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care 2023;46:e103-5.

Figures

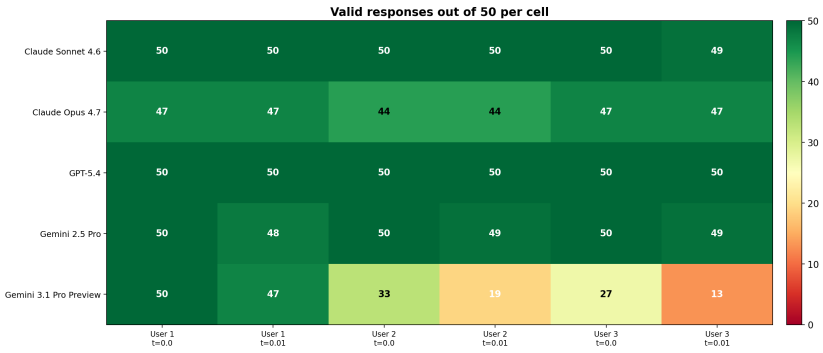

Figure 1. Response validity heatmap — valid responses out of 50 per (model, temperature, user) cell. Gemini 3.1 Pro Preview shows progressive truncation failure.

Figure 1. Response validity heatmap — valid responses out of 50 per (model, temperature, user) cell. Gemini 3.1 Pro Preview shows progressive truncation failure.

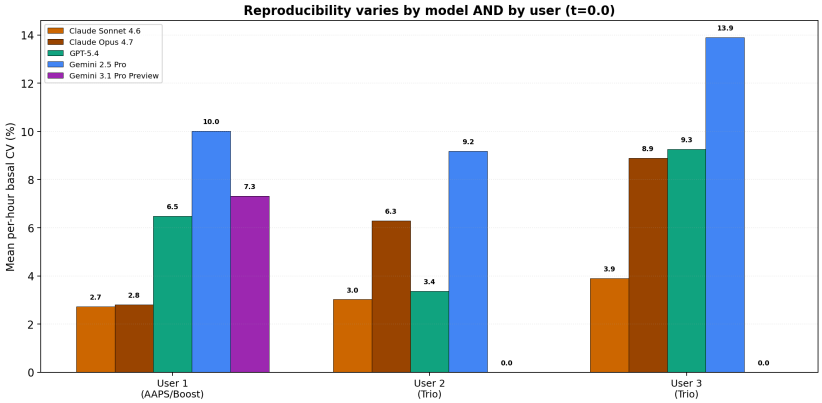

Figure 2. Mean per-hour basal CV at t=0.0 across three users. Sonnet consistently lowest; Gemini 2.5 Pro consistently highest.

Figure 2. Mean per-hour basal CV at t=0.0 across three users. Sonnet consistently lowest; Gemini 2.5 Pro consistently highest.

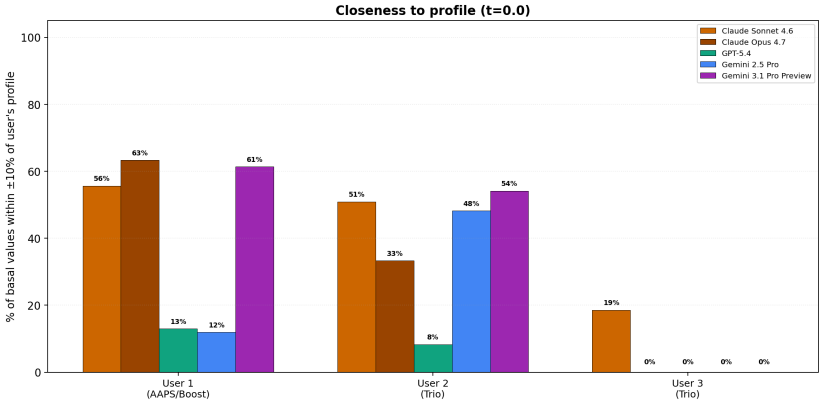

Figure 3. Closeness to profile — % of basal recommendations within ±10% of user's settings. All models collapse on User 3.

Figure 3. Closeness to profile — % of basal recommendations within ±10% of user's settings. All models collapse on User 3.

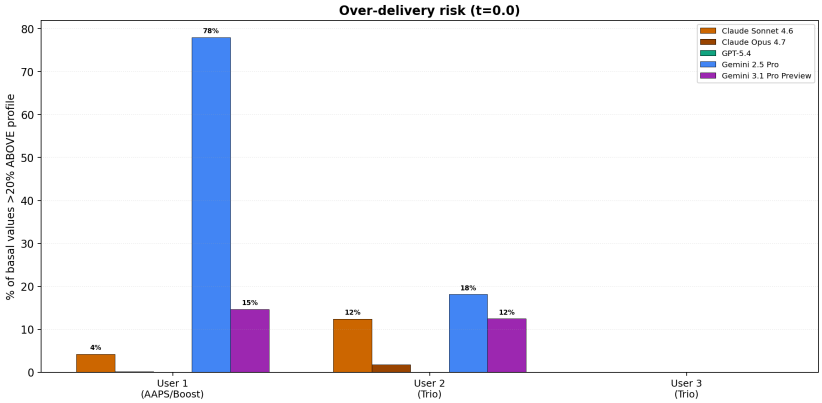

Figure 4. Basal recommendations significantly higher or lower than user's settings — % of values >20% above profile.

Figure 4. Basal recommendations significantly higher or lower than user's settings — % of values >20% above profile.

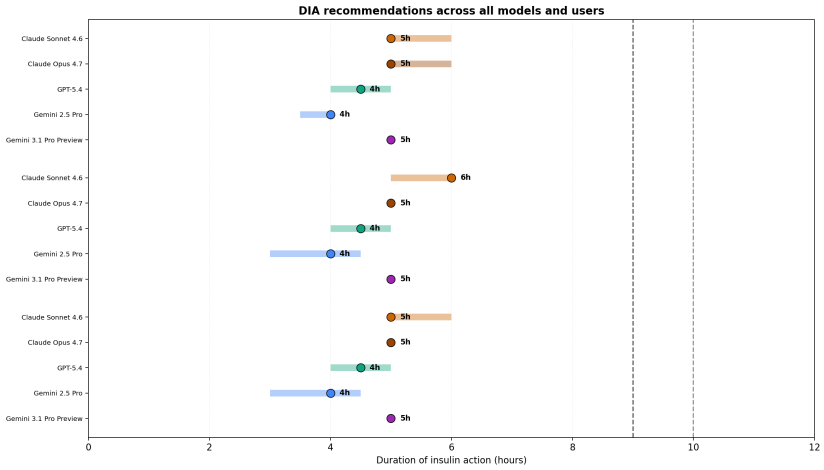

Figure 5. DIA recommendations vs user profiles. Every dot at 4-5 hours; every profile at 9-10 hours.

Figure 5. DIA recommendations vs user profiles. Every dot at 4-5 hours; every profile at 9-10 hours.

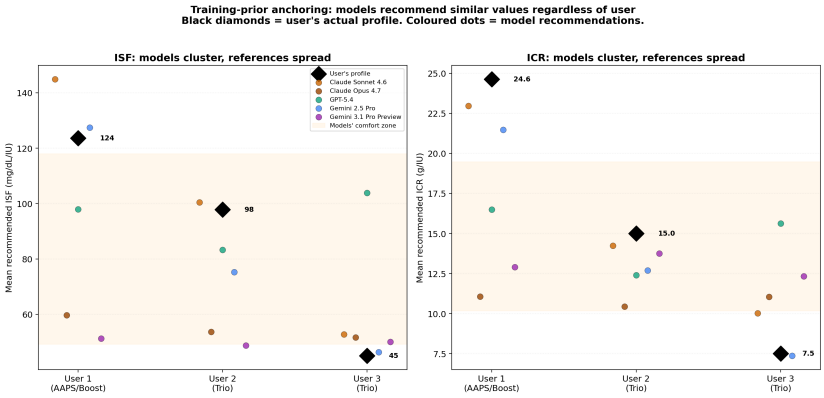

Figure 6. Training-prior anchoring: models recommend similar ISF and ICR regardless of user. Black diamonds = user's actual profile; coloured dots = model recommendations; orange zone = models' comfort zone. References spread wide while model recommendations cluster.

Figure 6. Training-prior anchoring: models recommend similar ISF and ICR regardless of user. Black diamonds = user's actual profile; coloured dots = model recommendations; orange zone = models' comfort zone. References spread wide while model recommendations cluster.

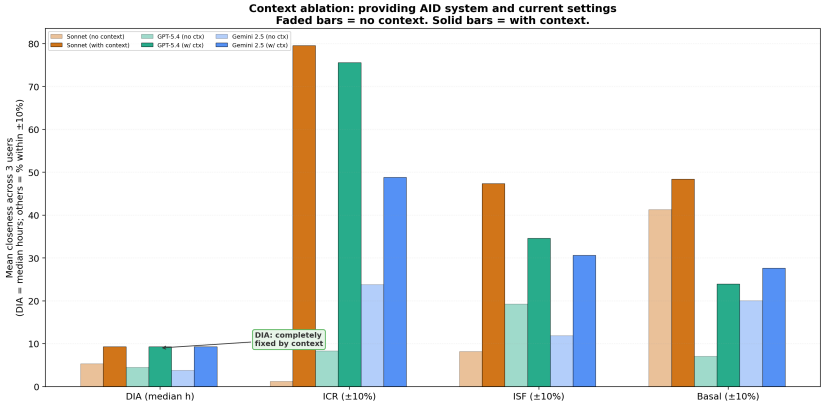

Figure 7. Context ablation: closeness-to-profile with (solid) and without (faded) AID system context and current settings. DIA completely fixed; ICR dramatically improved; ISF partially improved; basal mixed.

Figure 7. Context ablation: closeness-to-profile with (solid) and without (faded) AID system context and current settings. DIA completely fixed; ICR dramatically improved; ISF partially improved; basal mixed.

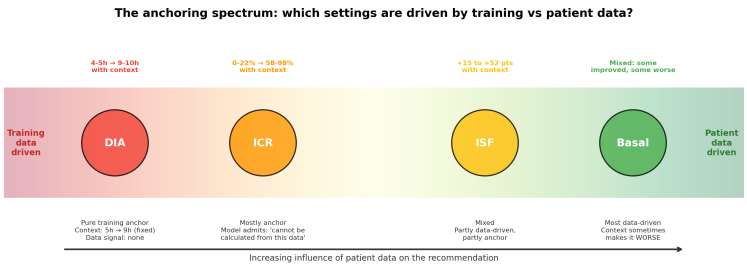

Figure 8. The anchoring spectrum: which settings are driven by training data vs patient data. DIA (pure training anchor) through ICR and ISF to basal (most data-driven).

Figure 8. The anchoring spectrum: which settings are driven by training data vs patient data. DIA (pure training anchor) through ICR and ISF to basal (most data-driven).

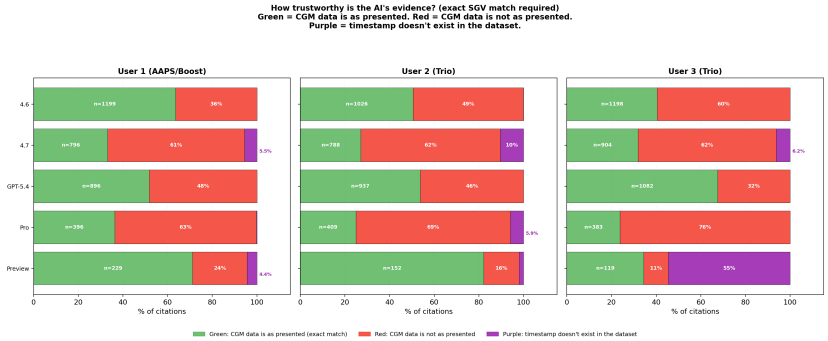

Figure 9. Deep citation verification — exact SGV match. Green = CGM data as presented; red = CGM data not as presented; purple = timestamp doesn't exist in the dataset.

Figure 9. Deep citation verification — exact SGV match. Green = CGM data as presented; red = CGM data not as presented; purple = timestamp doesn't exist in the dataset.

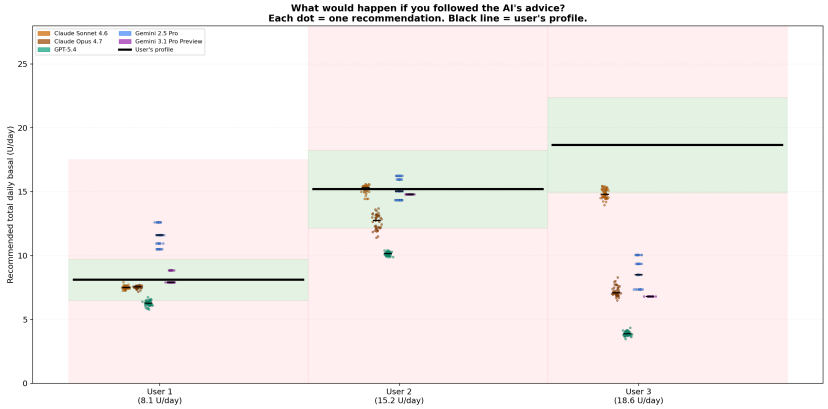

Figure 10. Total daily basal per run with ±20% danger zones.
Each dot = one recommendation; black line = user's profile.

Figure 10. Total daily basal per run with ±20% danger zones. Each dot = one recommendation; black line = user's profile.