Understanding Service Reliability of Large Language Models: An Empirical Characterization on Operator and User Reports

Yiren Bai

1st supervisor: Prof. dr. ir. Alexandru Iosup
daily supervisor: ir. Xiaoyu Chu
2nd reader: Prof. dr. ir. Tiziano De Matteis

Research Motivation

As the adoption of LLM services continues to scale, ensuring their reliability has become a critical challenge. However:

- Existing studies heavily rely on operator reports and lack access to large-scale, structured user feedback — **a lack of comprehensive failure data**.
- No effective tools exist to systematically collect, align, and compare operator and user reports — **an absence of standardized analytical approaches**.

These gaps hinder a comprehensive understanding and evaluation of LLM service reliability from both perspectives.

Research Questions

How to collect and understand LLM service reliability through different sources of failure data?

5 sub-questions (RQ):

- **RQ1:** How to **collect, process, and unify** operator- and user-reported **failure data** to support reliability analysis of LLM services?
- **RQ2:** How to characterize **failure recovery patterns** across LLM services?
- **RQ3:** What **temporal patterns** emerge in LLM service failures, and how can they inform proactive reliability strategies?
- **RQ4:** How strongly **correlated** are operator and user-reported failure signals, and what inter-service dependencies can the analysis identify?
- **RQ5:** To what extent do operator reports **align with** user experiences, and what does this reveal about provider reporting practices?

Research Contributions

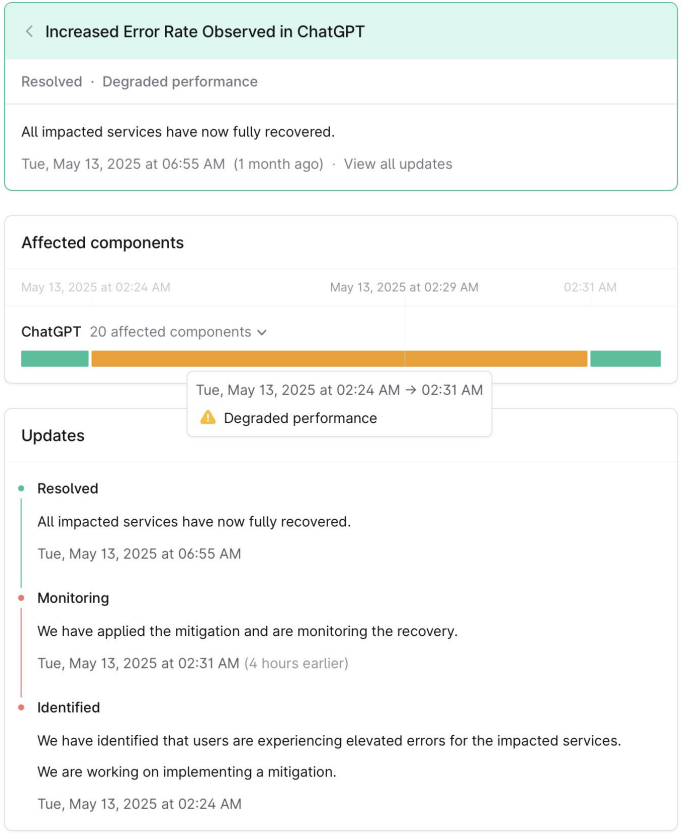
- **RC1 (Conceptual):** We propose a **data collection and analysis methodology** for understanding LLM service reliability through multiple sources of failure reports.
- **RC2 (Technical):** We implement a **data collection and analysis tool** for processing and characterizing LLM failure data.
- **RC3 (Conceptual):** We design **four types of failure analysis** on our collected datasets: failure-recovery modeling, temporal pattern analysis, correlation analysis, and consistency analysis. Each type targets a distinct dimension of service reliability.
- **RC4 (Technical):** We conduct **11 types of analysis** and summarize **28 important observations** to provide insights based on long-term failure data from prominent LLM services.
- **RC5 (Open Science):** We will release the collected datasets and the relevant toolkit as a contribution to open science.

Data Sources — Operator Reports

LLM services from 4 major providers:

- **OpenAI's ChatGPT**: https://status.openai.com
- **Anthropic's Claude**: https://status.anthropic.com
- **DeepSeek's DeepSeek**: https://status.deepseek.com
- **Character.AI**: https://status.character.ai

Data Sources — Detailed Incident Content



Increased Error Rate Observed in ChatGPT

Resolved · Degraded performance

All impacted services have now fully recovered.

Tue, May 13, 2025 at 06:55 AM (1 month ago) · View all updates

Affected components

May 13, 2025 at 02:24 AM — May 13, 2025 at 02:29 AM — 02:31 AM

ChatGPT 20 affected components

Tue, May 13, 2025 at 02:24 AM → 02:31 AM

Degraded performance

Updates

Resolved

All impacted services have now fully recovered.

Tue, May 13, 2025 at 06:55 AM

Monitoring

We have applied the mitigation and are monitoring the recovery.

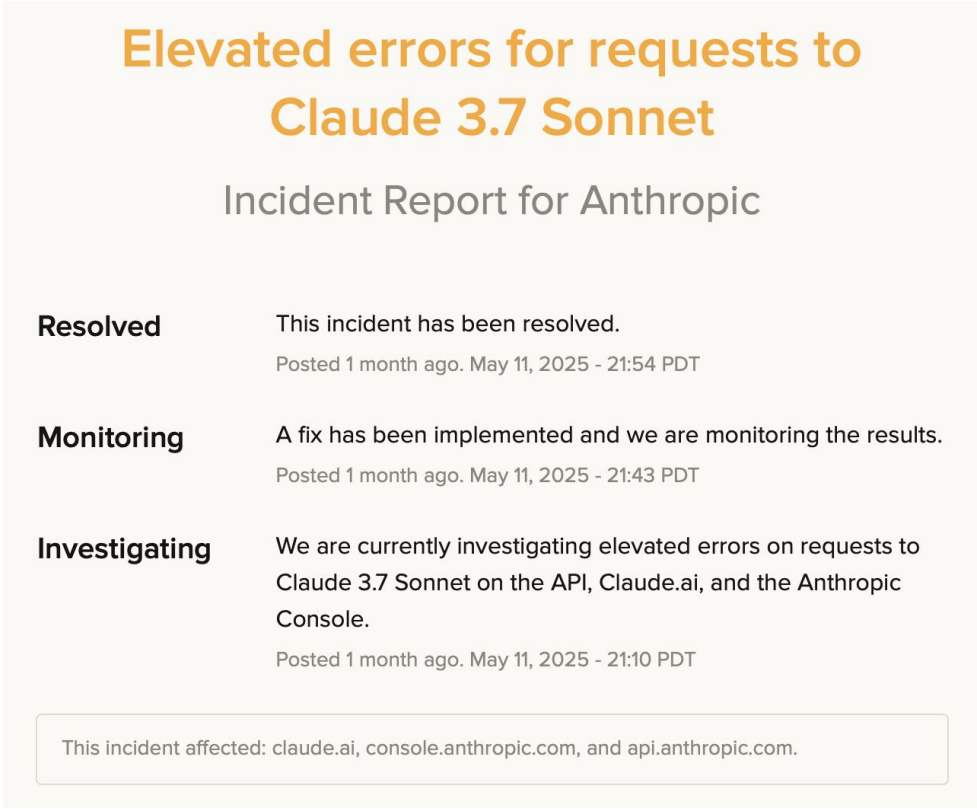
Tue, May 13, 2025 at 02:31 AM (4 hours earlier)

Identified

We have identified that users are experiencing elevated errors for the impacted services.

We are working on implementing a mitigation.

Tue, May 13, 2025 at 02:24 AM



Elevated errors for requests to Claude 3.7 Sonnet

Incident Report for Anthropic

Resolved This incident has been resolved.

Posted 1 month ago. May 11, 2025 - 21:54 PDT

Monitoring A fix has been implemented and we are monitoring the results.

Posted 1 month ago. May 11, 2025 - 21:43 PDT

Investigating We are currently investigating elevated errors on requests to Claude 3.7 Sonnet on the API, Claude.ai, and the Anthropic Console.

Posted 1 month ago. May 11, 2025 - 21:10 PDT

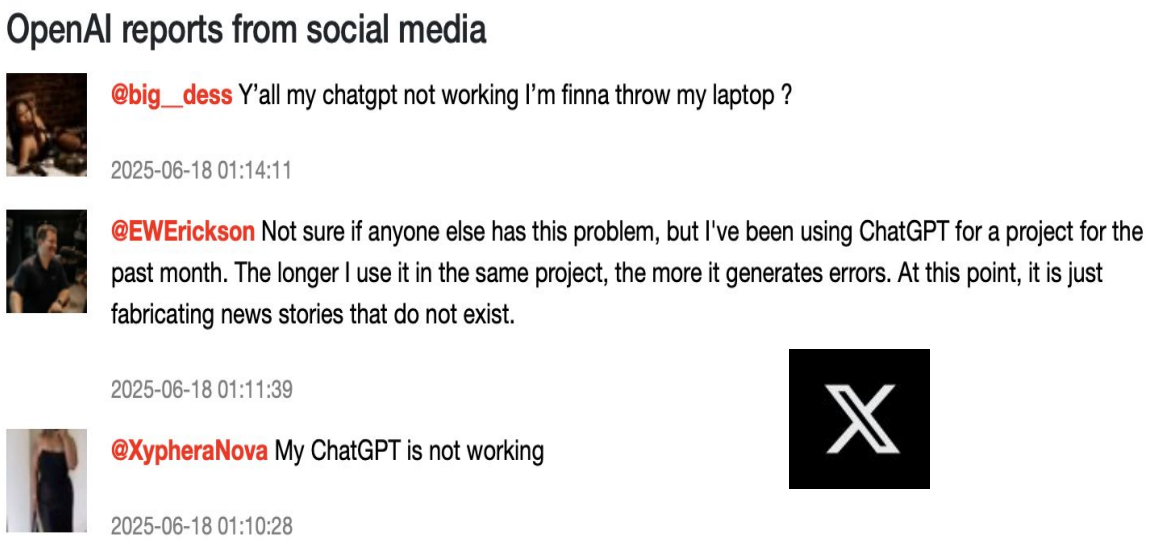
This incident affected: claude.ai, console.anthropic.com, and api.anthropic.com.

Data Sources — User Reports

Data Sources — Downdetector & X/Twitter

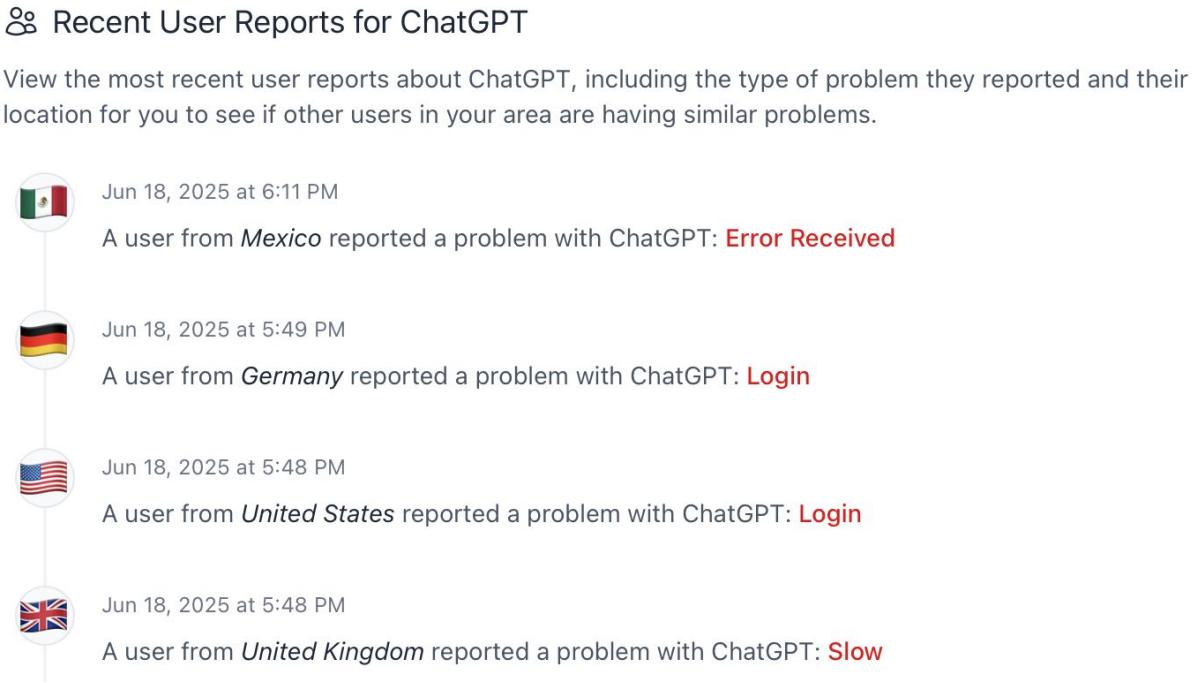
- E.g. https://downdetector.com/status/openai/

Comments:

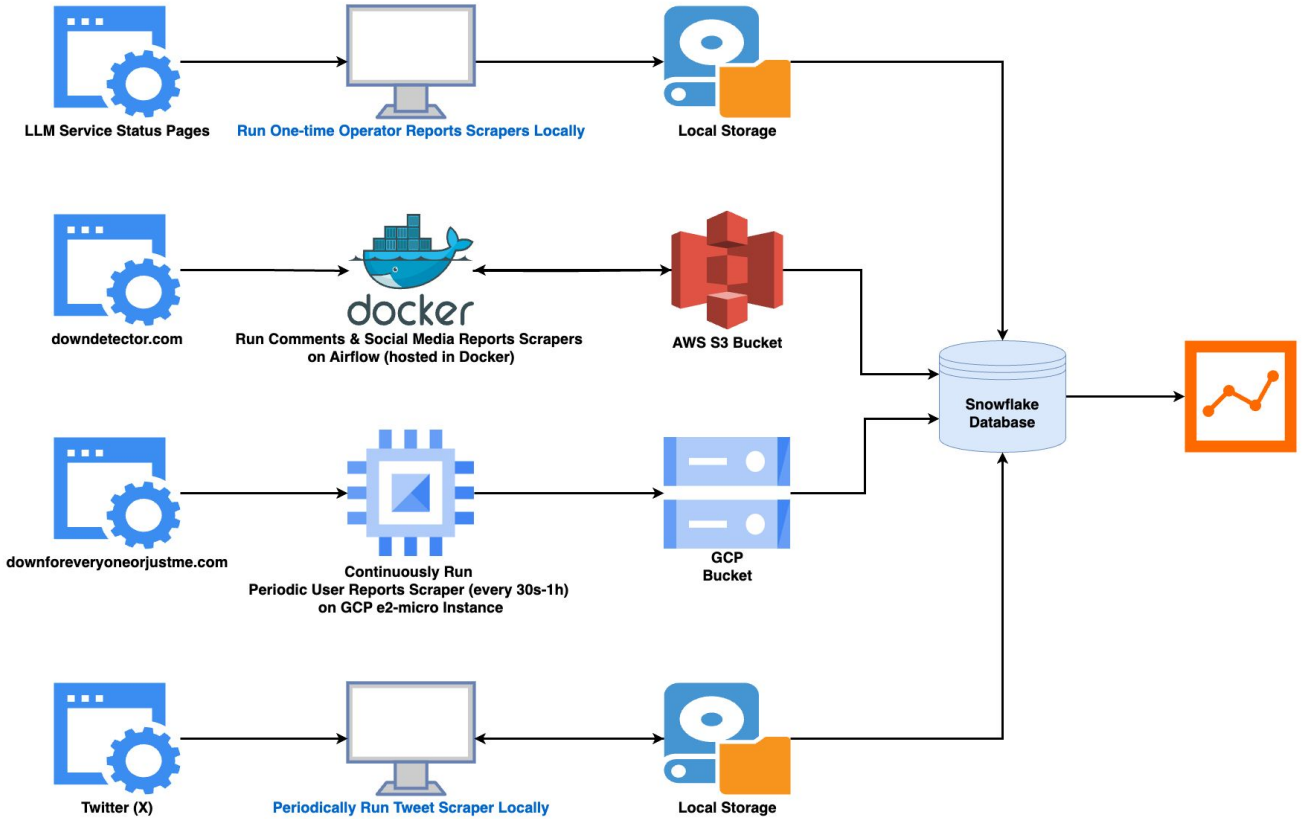


Data Sources — Down for Everyone or Just Me

- E.g. https://downforeveryoneorjustme.com/chatgpt



RQ1: Data collection pipeline



RQ1: Summary of collected datasets

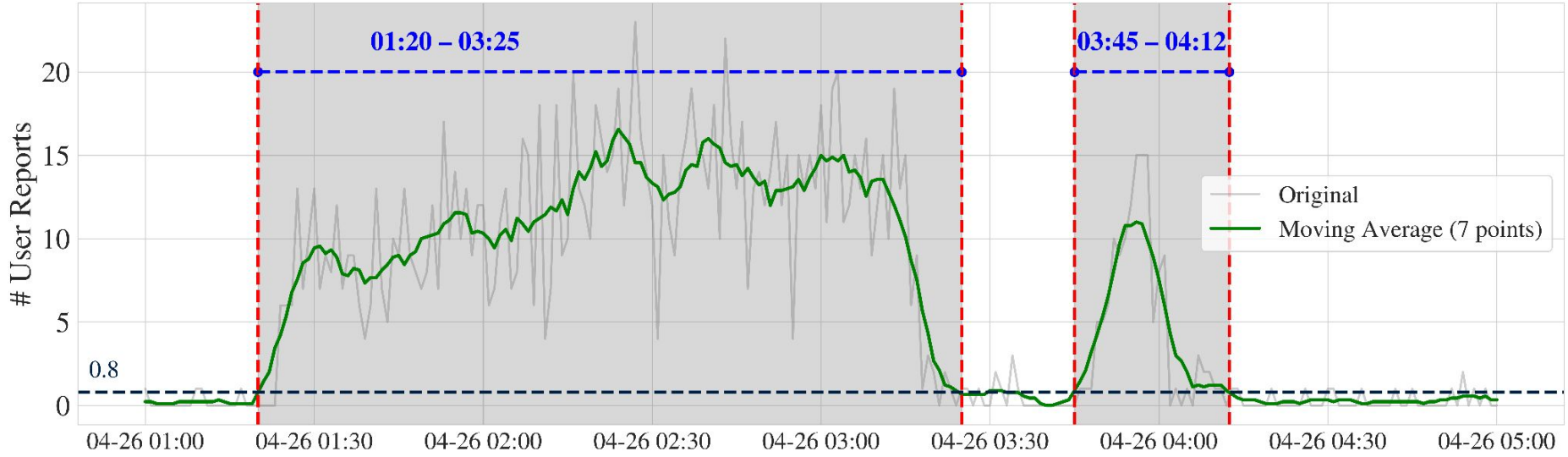
Platform	Dataset Content	Service	First Date	Last Date	Coverage [D]	# of Reports
Status Pages	Incident Summary	ChatGPT	2023-12-01	2025-05-18	535	179
		Claude	2023-12-01	2025-05-18	535	227
		DeepSeek	2024-05-01	2025-05-18	383	37
		Character.AI	2023-12-01	2025-05-18	535	87
	Incident Detail	ChatGPT	2023-12-01	2025-05-18	535	179
		Claude	2023-12-01	2025-05-18	535	227
		DeepSeek	2024-05-01	2025-05-18	383	37
		Character.AI	2023-12-01	2025-05-18	535	87
DownDetector	Comments	ChatGPT	2025-01-02	2025-05-18	137	2,551
		Claude	2025-01-30	2025-05-18	109	18
		DeepSeek	2025-01-29	2025-05-18	110	57
		Character.AI	2025-01-04	2025-05-18	135	12,237
	Social Media Reports	ChatGPT	2025-02-11	2025-05-18	97	2,966
		Claude	2024-12-29	2025-05-18	141	1,006
		DeepSeek	2025-02-18	2025-05-18	90	406
		Character.AI	2024-10-24	2025-05-18	207	47
DownForEveryoneOrJustMe	User-Reported Issues	ChatGPT	2025-02-13	2025-05-18	95	38,160
		Claude	2025-02-06	2025-05-18	102	1,900
		DeepSeek	2025-02-12	2025-05-18	96	1,377
		Character.AI	2025-02-13	2025-05-18	95	73,285
Twitter/X	Tweets	ChatGPT	2025-01-01	2025-05-18	138	10,740
		Claude	2025-01-01	2025-05-18	138	8,205
		DeepSeek	2025-01-01	2025-05-18	138	6,000
		Character.AI	2025-01-01	2025-05-18	138	5,461

RQ1: Data processing

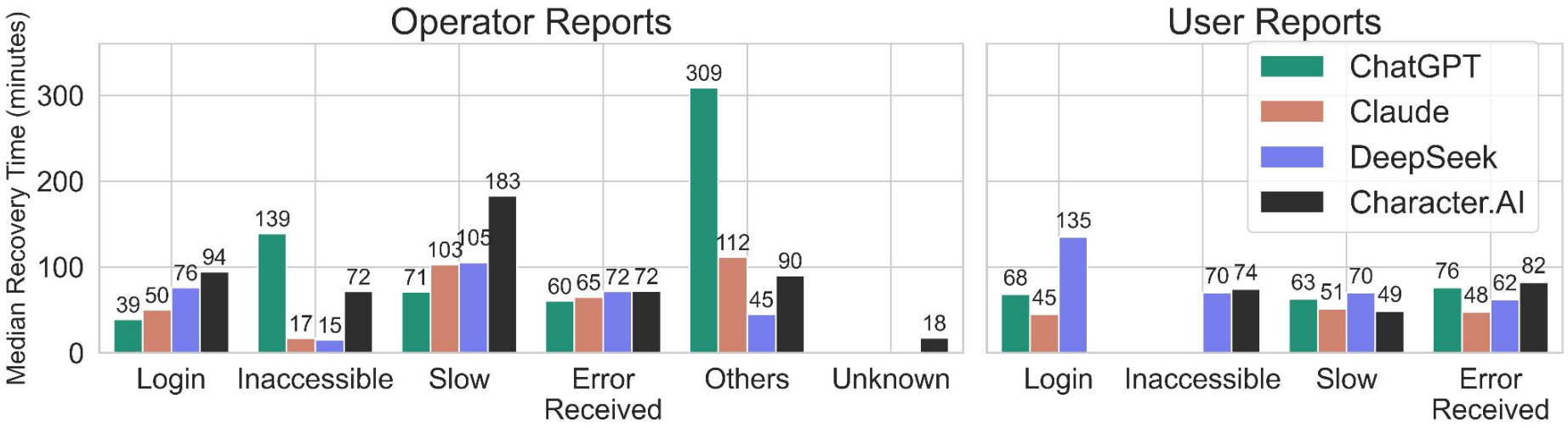
Three main steps:

- **Convert timestamps** shown in different time zones **to UTC**
- **Extract failure types from free text**, such as user comments and incident descriptions, **using few-shot prompts** via the OpenAI API
  - Failure Types: ***Login, Inaccessible, Slow, Error Received, Others, Unknown***
- **Infer user-reported failure periods** by aggregating user reports into 1-minute bins and **applying a 7-minute moving average** to the time series of report counts

RQ1: Data processing — Inferring user-reported failure periods

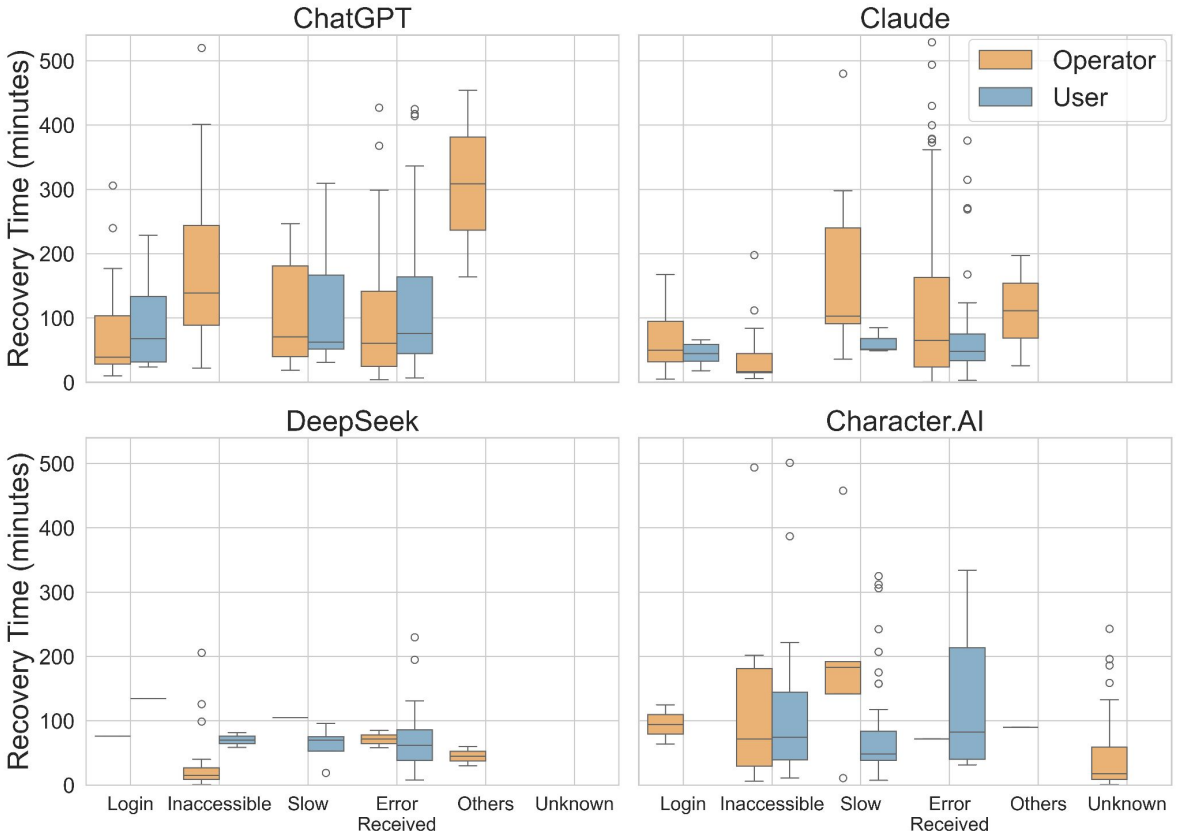


RQ2: Failure recovery patterns — Impact of failure type on median recovery time



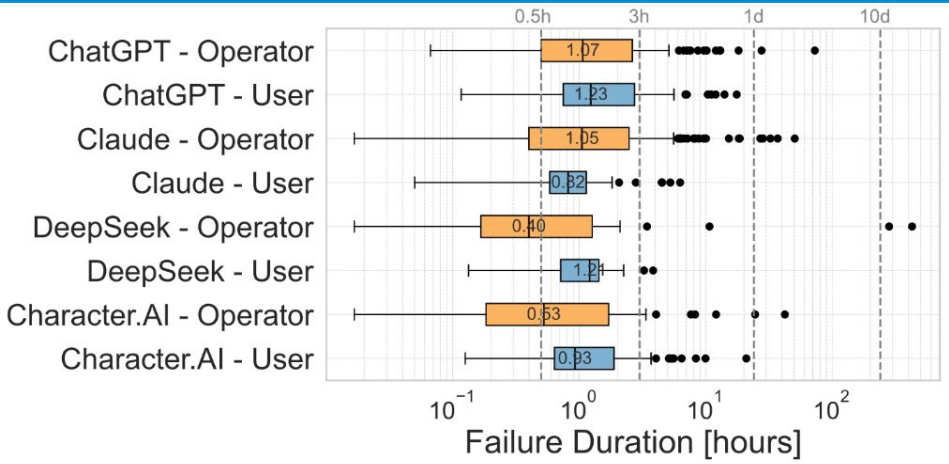
- **O1:** In **operator** reports, ***Slow*** failures generally have the **longest median recovery durations** compared to ***Login***, ***Inaccessible***, and ***Error Received*** failures.
- **O2:** **Operator** reports exhibit **greater variability** in median recovery durations, **both within and across failure types** (average within-type variance: 1585; across-type: 1839). In contrast, **user** reports show **much lower variability** (within-type: 631; across-type: 377).

RQ2: Failure recovery patterns — Recovery time distributions

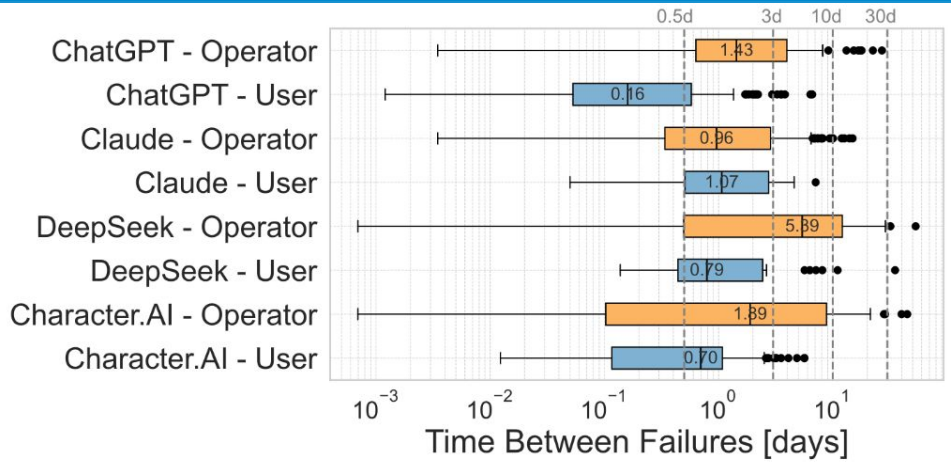


- **O3:** **Claude** exhibits **larger discrepancies** between operator and user **recovery time distributions** across failure types (average IQR: 117 vs. 29 minutes), whereas **ChatGPT** displays **more consistent** patterns between the two sources (111 vs. 112 minutes, respectively).

RQ2: Failure recovery patterns — Distributions of MTTR and MTBF



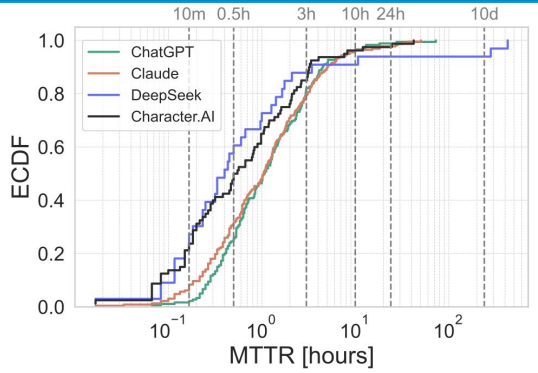
(a) MTTR distributions across LLM services.



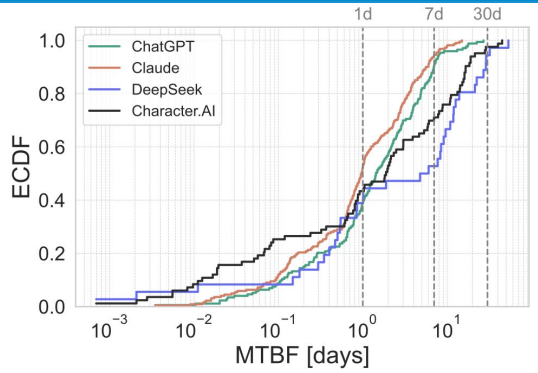
(b) MTBF distributions across LLM services.

- **O4:** **ChatGPT** has the **slowest recovery** from failures, with the highest median MTTR reported by both the **operator** (1.07 hours) and the **user** (1.23 hours).
- **O5:** **DeepSeek** is the **most reliable** service on the **operator side**, with the lowest median MTTR (0.40 hours) and the highest median MTBF (5.39 days), whereas **Claude ranks highest** in **user-side reliability** (median MTTR: 0.82 hours; median MTBF: 1.07 days).

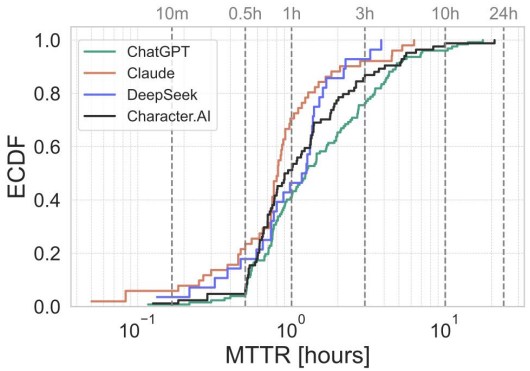
RQ2: Failure recovery patterns — ECDFs for MTTR and MTBF



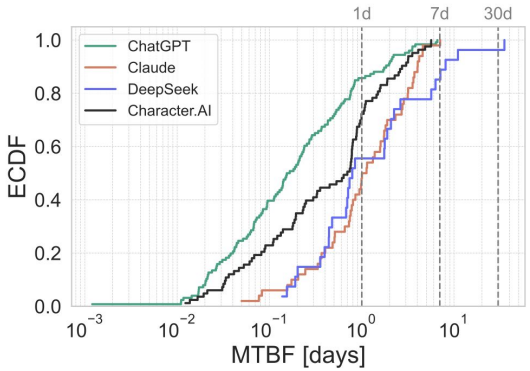
(a) MTTR - Operator



(b) MTBF - Operator



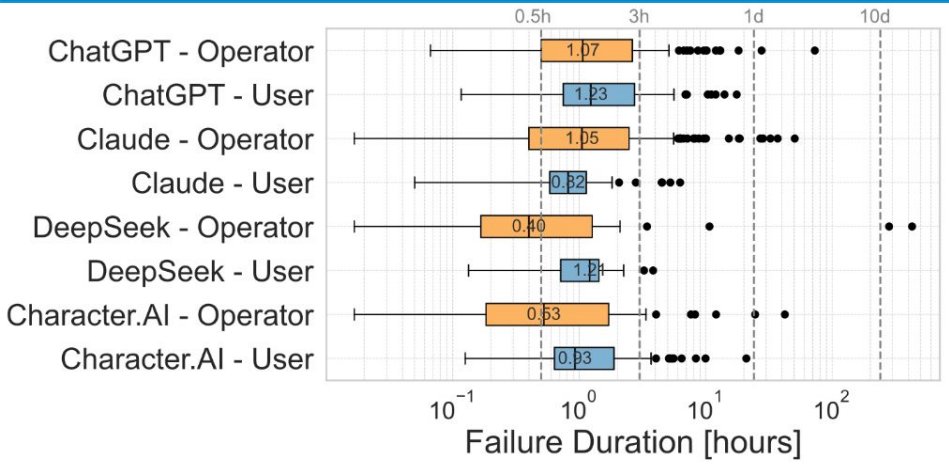
(c) MTTR - User



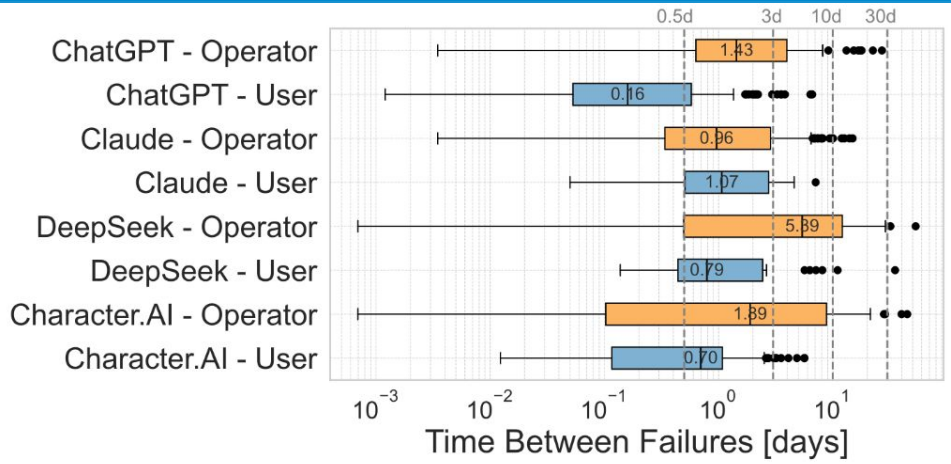
(d) MTBF - User

- **O6:** On the **operator** side, **DeepSeek resolves failures most quickly** (60.61% of durations ≤ 0.5 hours) and has the **longest failure intervals** (47.22% of intervals > 7 days).
- **O7:** From the **user** perspective, **Claude** exhibits the **fastest recovery** (70.59% of durations ≤ 1 hour) and the **longest failure intervals** (52.00% of intervals > 1 day).

RQ2: Failure recovery patterns — Distributions of MTTR and MTBF



(a) MTTR distributions across LLM services.



(b) MTBF distributions across LLM services.

- **O8:** **DeepSeek** exhibits the **largest discrepancy in median MTTR** between operator and user reports (0.81-hour gap), whereas **ChatGPT** demonstrates the **highest alignment** (0.16-hour gap).
- **O9:** **DeepSeek** also exhibits the **largest divergence in median MTBF** between operator and user reports (4.6-day gap), whereas **Claude** shows the **highest consistency** (0.11-day gap).

RQ2: Failure recovery patterns — Comparative analysis across services

Table 3.1: Comparative failure metrics of LLM services based on operator reports.

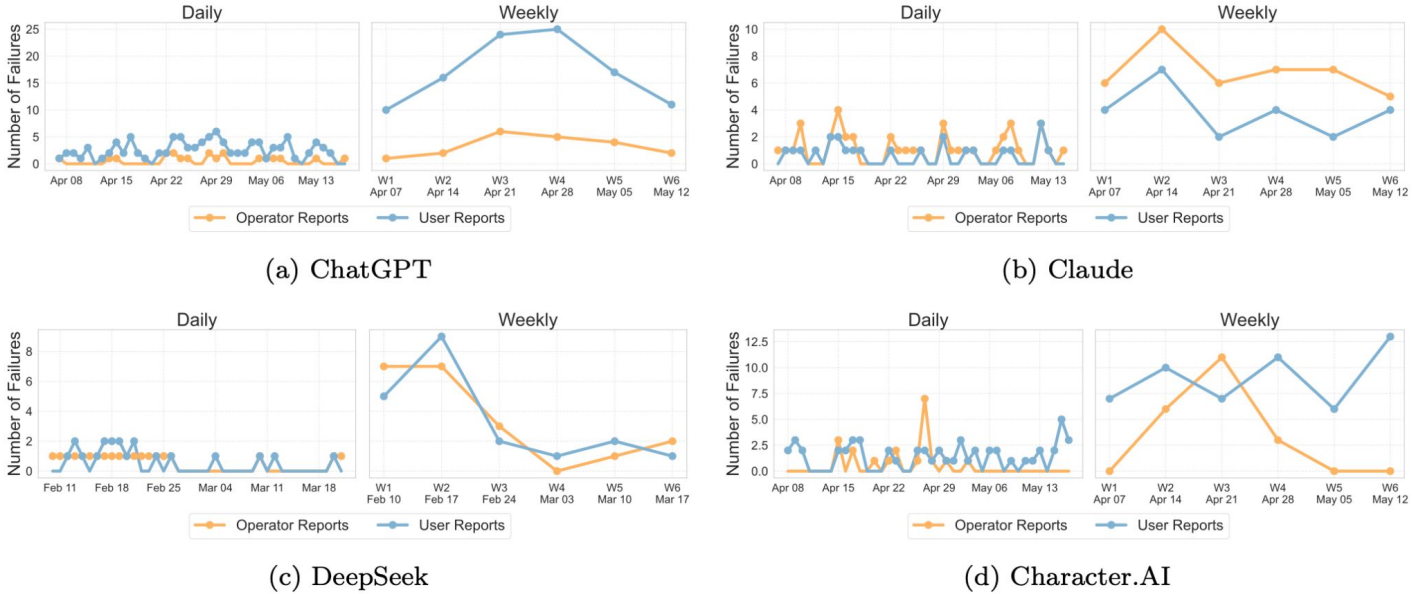
Metric	ChatGPT	Claude	DeepSeek	Character.AI
Mean Time to Recovery (MTTR) [h]	2.504	2.600	22.384	**2.056**
Median Failure Duration (MFD) [h]	1.067	1.050	**0.400**	0.525
Tail [P95] Failure Duration (TFD) [h]	**7.677**	8.163	118.927	**7.663**
Mean Time Between Failures (Mean MTBF) [d]	2.951	2.069	**9.370**	6.076
Median Time Between Failures (Median MTBF) [d]	1.431	0.960	**5.393**	1.885
Tail [P95] Time Between Failures (TTBF) [d]	8.039	7.413	**29.666**	21.064
Failure Frequency [failures/day]	0.338	0.471	**0.099**	0.165
Failure Type Entropy [bits]	**1.444**	1.029	1.165	1.400
Availability Percentage [%]	96.472	94.899	91.763	**98.699**
Avg. Failure Impact Level (Ordinal)	**1.458**	1.584	2.600	1.942

Table 3.2: Comparative failure metrics of LLM services based on user reports.

Metric	ChatGPT	Claude	DeepSeek	Character.AI
Mean Time to Recovery (MTTR) [h]	2.285	**1.179**	1.230	1.869
Median Failure Duration (MFD) [h]	1.233	**0.817**	1.208	0.925
Tail [P95] Failure Duration (TFD) [h]	6.938	4.500	**2.899**	5.548
Mean Time Between Failures (Mean MTBF) [d]	0.584	1.697	**3.438**	0.981
Median Time Between Failures (Median MTBF) [d]	0.160	**1.066**	0.789	0.698
Tail [P95] Time Between Failures (TTBF) [d]	2.757	4.396	**10.135**	3.491
Failure Frequency [failures/day]	1.426	0.552	**0.293**	0.951
Failure Type Entropy [bits]	0.742	0.978	**1.234**	**1.234**
Availability Percentage [%]	86.421	97.291	**98.497**	92.597

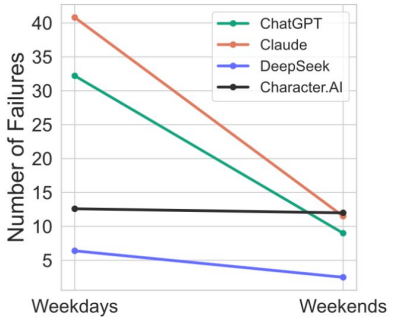
- **O10:** In terms of **failure frequency** and service **availability**, **Character.AI** is the **most reliable** service from the **operator** perspective (0.165/day, 98.699%), whereas **DeepSeek ranks highest** on the **user** side (0.293/day, 98.497%).

RQ3: Temporal patterns of failures — Time series of daily and weekly failure counts

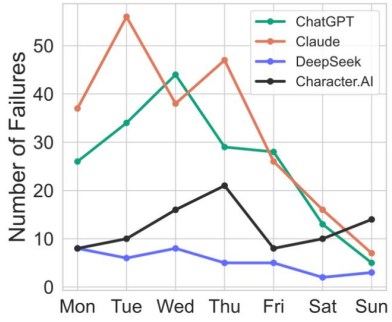


- **O11:** **ChatGPT** exhibits the most consistent gap between operator and user **failure counts**, with **user** reports **exceeding operator** reports on nearly all days and every week.
- **O12:** **Claude** shows the opposite pattern: **operator**-reported failures **frequently outnumber user** reports, particularly in weekly aggregates.

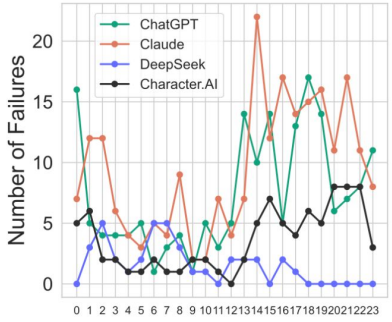
RQ3: Temporal patterns — Temporal distribution of failures



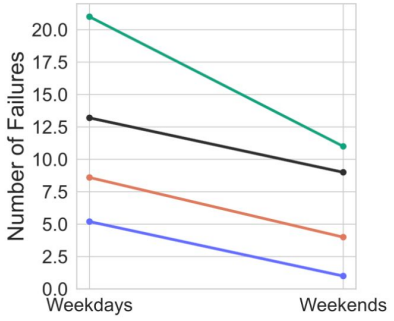
(a) Weekday vs. Weekend (Operator)



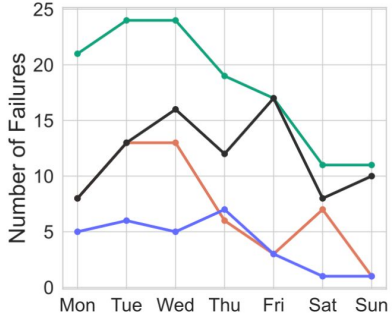
(b) Day of Week (Operator)



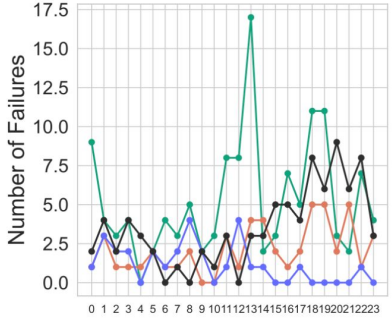
(c) Hour of Day (Operator)



(d) Weekday vs. Weekend (User)



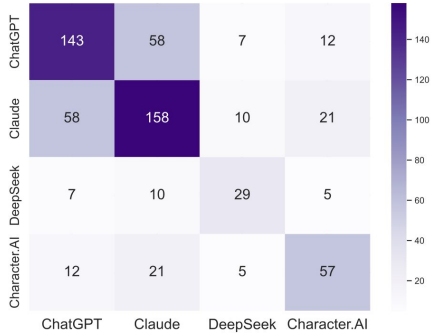
(e) Day of Week (User)



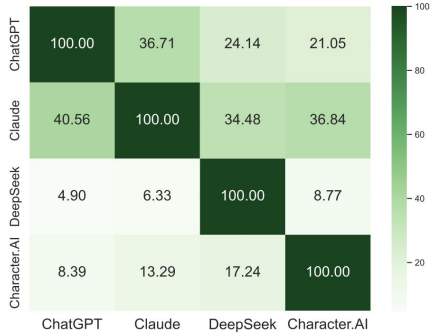
(f) Hour of Day (User)

- **O13:** Both **operator**- and **user**-reported failures occur **more frequently** on **weekdays** than on weekends.
- **O14:** All services show **peak** failures during their respective **regional working hours**: DeepSeek between 01:00 and 10:00 UTC (**09:00–18:00 CST**), and the others between 13:00 and 00:00 UTC (**06:00–17:00 PDT**).

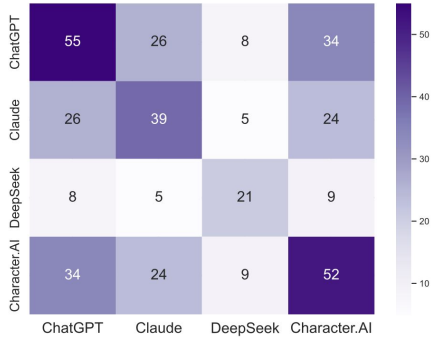
RQ4: Cross-service correlations — Failure co-occurrence



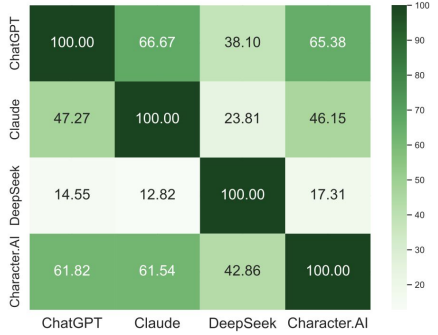
(a) Co-occurrence failures in days count based on operator reports.



(b) Conditional probabilities (%) of co-occurrence failures based on operator reports. Notes: y-axis = service A, x-axis = service B, cells = $P(A \mid B)$.



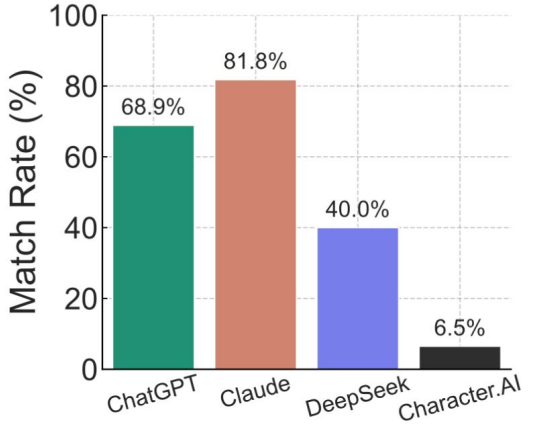
(c) Co-occurrence failures in days count based on user reports.



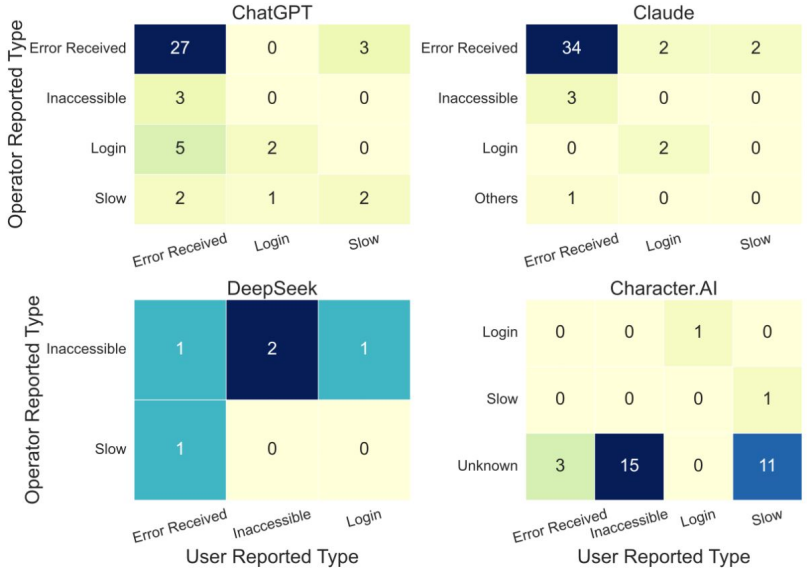
(d) Conditional probabilities (%) of co-occurrence failures based on user reports. Notes: y-axis = service A, x-axis = service B, cells = $P(A \mid B)$.

- **O15:** On the **operator** side, **failure co-occurrence** is **highest** between **ChatGPT and Claude**, which exhibit elevated conditional probabilities of 40.56% and 36.71%, respectively. In contrast, **DeepSeek** shows **minimal co-occurrence** with other services, indicating a high degree of operational independence.
- **O16:** **User**-reported failures reveal **strong cross-service co-occurrence**, especially between **ChatGPT and Character.AI**, which exhibit mutual conditional probabilities exceeding 60%.

RQ5: Consistency analysis — Match rates of failure types between the two sources



(a) Bar chart of match rates for each LLM service, reflecting the proportion of operator-reported failures whose failure types match the most frequently reported user failure types in the same time window.



(b) Heatmaps showing the joint distribution of operator- and user-reported failure types for each LLM service.

- **O18:** The **most frequent mismatch** occurs when operator-reported failures are misclassified by users as ***Error Received***, which accounts for an average of 62.7% of all misclassified cases across services.

- **O17:** **Claude** exhibits the **highest failure type match rate** (81.8%), whereas **Character.AI** shows the **lowest** match rate (6.5%).

RQ5: Consistency of failure periods between the two sources

Table 6.1: Consistency-related metrics summarizing the overlap and timing alignment between operator and user reports. Legend: m = minute(s).

Metric	ChatGPT	Claude	DeepSeek	Character.AI
# of Operator Reports	45	74	4	40
# of User Reports	127	50	12	83
# of Overlapping Reports	37	37	2	20
Coverage of Operator [%]	82.22	50.00	50.00	50.00
Coverage of User [%]	29.13	74.00	16.67	24.10
Mean User Lead Time [m]	72.86	-1.12	14.71	78.13
Median User Lead Time [m]	31.55	3.00	14.71	43.50

- **O19:** **ChatGPT** and **Claude** show relatively **strong alignment** between operator and user reports, though for different reasons: **ChatGPT** has a **high operator coverage** of 82.22%, while **Claude** shows **accurate user reporting** at 74.00%, along with **closely aligned user detection**, with a median user lead time of 3 minutes.

Summary

Take home messages:

- **ChatGPT recovers the slowest** across both sources; **DeepSeek** has the **longest failure intervals** on the **operator** side, while **Claude** shows the **longest** on the **user** side.
- **DeepSeek** exhibits the **largest discrepancy** between operator and user reports, in terms of both **median recovery durations** and **median failure intervals**.
- Both operator- and user-reported failures display clear **periodic and diurnal patterns across services**, occurring more frequently on weekdays and peaking during the working hours of each service's primary user region.
- **ChatGPT** and **Claude** show relatively **strong alignment** between operator and user reports in terms of **failure periods**, though for different reasons: ChatGPT due to high operator coverage, and Claude due to accurate user reporting.

Thank you! Welcome to ask any questions.