Picker

September 2026

# Lost in translation? How AI systems describe patients' experiences of care

Authors: Jenny King & Molly Hopson

# Picker

Picker is a leading international health and social care charity. We carry out research to understand individuals' needs and their experiences of care. We are here to:

- Influence policy and practice so that health and social care systems are always centred around people's needs and preferences.
- Inspire the delivery of the highest quality care, developing tools and services which enable all experiences to be better understood.
- Empower those working in health and social care to improve experiences by effectively measuring, and acting upon, people's feedback.



Published by and available from:

Picker Institute Europe
Suite 6, Fountain House,
1200 Parkway Court,
John Smith Drive,
Oxford OX4 2JY
Tel: 01865 208100
Email: info@pickereurope.ac.uk
Website: picker.org

Registered Charity in England and Wales: 1081688
Registered Charity in Scotland: SC045048
Company Limited by Registered Guarantee No 3908160

Picker Institute Europe has UKAS accredited certification for ISO20252:2019 (GB08/74322) via SGS, and ISO27001:2022 and ISO 27701:2019 (23715) via Alcumus ISOQAR. We comply with Data Protection Laws including the General Data Protection Regulation, the Data Protection Act 2018 and the Market Research Society's (MRS) Code of Conduct.

# Contents

# Key messages

This report examines how four generative Artificial Intelligence (AI) systems – OpenAI's ChatGPT, Microsoft's Copilot, Google's Gemini, and Anthropic's Claude – represented evidence about the experiences of adults receiving NHS acute inpatient care in England. Using findings from the 2024 NHS Adult Inpatient Survey (1) as a comparator, we explored how AI-generated responses reflected, interpreted, and framed patient experience evidence.

### AI systems broadly reproduced the main quantitative findings of the NHS Adult Inpatient Survey

All four AI systems gave a broadly accurate summary of the main findings from the 2024 NHS Adult Inpatient Survey. They consistently identified modest improvements in people's experiences with staff while highlighting ongoing challenges around access to care and discharge. They also reflected the main differences between patient groups: particularly those related to age, frailty, disability, and how people were admitted to hospital. Overall, this suggests that generative AI can provide a useful overview of headline findings from patient experience surveys.

### AI systems interpreted the evidence, rather than simply summarising it

The AI systems did more than report the survey findings. They often presented findings within wider explanatory narratives relating to issues and policy priorities such as pressure on NHS services, recovery after the COVID-19 pandemic, and how services work together. These interpretations reflected both the choices made by the AI systems and the way the evidence had already been framed in reports, press releases, and other published sources. As a result, AI-generated responses should be seen as a further stage in the translation of patient experience evidence rather than as straightforward summaries of what patients reported.

### The evidence was often the same, but the story was different

The four AI systems often used the same sources but described patient experience in different ways. The differences came less from the source evidence they used and more from the 'choices' they made about what to include and what to emphasise. This means that people asking the same question to different AI systems may receive different accounts of patient experience.

### Some aspects of patient experience received more attention than others

Staff interactions, access to care, and discharge were consistently given the most attention across all four AI systems. In contrast, the physical environment of care and several findings about specific patient groups were mentioned less often or left out altogether. However, the AI systems were summarising evidence that had already been synthesised, meaning these patterns may partly reflect the emphasis and structure of the

source summaries as well as the AI systems' own prioritisation. As a result, AI-generated summaries may give greater prominence to some aspects of patient experience than others.

### Qualitative evidence was less faithfully preserved

Quantitative (numeric) findings from the survey were generally reported accurately, whereas qualitative evidence (such as people's comments) was more likely to be summarised and interpreted. In one instance, themes from the qualitative report appeared to be reconstructed as 'patient quotations' that were not present in the source material. This made it harder to distinguish patient voice from AI-generated interpretation.

### Findings about inequalities were the least consistent

The AI systems were less consistent when describing differences between patient groups than when summarising the overall survey findings or discharge experiences. Although they generally identified the main patterns, they were more likely to add explanations, describe groups differently, or include details that could not be clearly traced back to the published survey. This may reflect the fact that subgroup findings are more complex and are often presented in tables and charts rather than as a narrative account, requiring greater extrapolation and interpretation compared to the headline findings.

### Person centred perspectives were largely maintained

Despite differences in emphasis and interpretation, all four AI systems described care from the patient's perspective. When they discussed wider issues such as NHS performance, service pressures or policy, these were usually explained in terms of how they affected patients rather than replacing the patient perspective with a system focused one. This suggests that the person centred perspective of the original evidence was generally preserved.

### Overall message

Generative AI can provide a useful overview of patient experience evidence, particularly when summarising the main findings. However, AI-generated responses are not neutral summaries. They involve 'choices' about what to highlight, what to leave out, and how to explain the evidence. As is often the case with generative AI systems, the reasons why specific pieces of evidence are foregrounded can be opaque.

This matters for anyone using AI to understand patient experience. Researchers, healthcare leaders, patient experience teams, policymakers, and members of the public may receive different accounts of the same evidence depending on which AI system they use (or, indeed, how they choose to present their questions – although this is beyond the scope of our current investigation). It may be difficult to tell where the published evidence ends, and the AI's interpretation begins.

While headline survey findings were generally represented accurately, qualitative evidence and findings about differences between patient groups were more likely to be interpreted and reframed. As a result, different AI systems may highlight different aspects of patient experience and inequalities, even when drawing on the same evidence.

Given the large volumes of patient experience information now available, it is easy to envisage scenarios where users may find it useful to turn to generative AI models to assist them in developing summaries. When decisions depend on a detailed, authentic, understanding of patient experience, AI-generated summaries should be used alongside the original evidence rather than as a replacement for it.

# Introduction and context

Generative Artificial Intelligence (AI) tools are increasingly being used to answer questions about healthcare, bring together information from different sources, and explain complex topics (2) . A policymaker, analyst, clinician, or member of the public can ask an AI system what patients think about hospital care and receive an answer within seconds. However, we still know relatively little about how these systems represent patient experience evidence, what they emphasise, and what they leave out[1].

This matters because patient experience is recognised as a core component of healthcare quality, alongside safety and clinical effectiveness (3). Findings from patient experience surveys are used to improve services and inform regulation and policy across healthcare systems (4). As more people turn to generative AI to understand healthcare evidence, the way these AI systems present patient experience findings is likely to influence how healthcare quality is understood.

Every summary of evidence involves choices about what to include, what to leave out and what to emphasise. Generative AI is no different. The difference is that these choices are made very quickly and with little transparency: although terms like 'reasoning' are commonly used to describe the means by which AI systems arrive at their outputs, the algorithmic process is opaque and not comparable to human decision making. Nevertheless, choices are a feature of outputs – because rather than just directing people to the original evidence, AI systems produce their own summaries. In doing so, they shape how that evidence is understood.

This is particularly important for patient experience evidence. Unlike many other measures of healthcare quality, patient experience evidence is intended to reflect the perspectives of users. The overall picture depends not only on what patients reported, but also on which experiences are highlighted and which patients' experiences receive the most attention.

There are concerns that large language models (LLMs) may overemphasise some findings, overlook others, or change the way evidence is presented (5). In healthcare, this could produce summaries that lose some of the balance or complexity of the original evidence (6). AI systems may place greater emphasis on patterns that are more common in their training data or source material, while giving less attention to experiences and needs that are less well represented. This could make inequalities in care less visible and lead to an incomplete picture of healthcare quality (7).

---

[1] For readability, this report sometimes uses language related to thinking and decision-making – like "choose", "interpret", or "emphasise" – when describing AI-generated outputs. These terms are intended to offer shorthand descriptions of analogues between AI-generated outputs and the human processes that would be required to create similar outputs. They should not be taken to imply human-like reasoning, intention, understanding, or judgement in AI systems.

Although interest in generative AI in healthcare is growing, little attention has been paid to how these AI systems represent patients' reported experiences. This matters more as AI use becomes integrated into routine practice. Recent UK survey data suggest that AI use is now widespread, with almost three-quarters of adults reporting that they had used AI in the previous month (8). Within the National Health Service (NHS) in England, AI-powered assistants are already being used to help staff find information, complete administrative tasks, and support decision-making (9).

As these tools become more widely used, people may accept AI-generated responses as accurate summaries of the evidence. It is therefore important to understand whether these summaries accurately reflect what patients reported, and where they differ from the original evidence.

## Aim and areas of focus

This report examines what happens when patient experience evidence is turned into AI-generated summaries. Using findings from the 2024 NHS Adult Inpatient Survey, it compares AI-generated responses with the original quantitative and qualitative evidence.

The report explores:

- How closely AI-generated accounts reflect patients' reported experiences.
- Which aspects of inpatient care are given the greatest prominence in AI-generated summaries.
- How AI systems frame and interpret patient experience evidence.
- Which experiences, patient groups, or findings are left out or underrepresented.

By identifying where AI summaries reflect the evidence, where they reinterpret it, and where important information is lost, the report provides new evidence about how generative AI shapes people's understanding of patient experience.

This study examines AI's representation of patient experience evidence and does not aim to compare chatbot performance.

# Methods

## AI systems

Four publicly available generative AI systems were included:

- ChatGPT (GPT-5.5)
- Claude (Sonnet 4.6)
- Google Gemini (3.5 Flash)
- Microsoft Copilot (Smart. No public internal code name)

These AI systems were selected as they are widely used and available to the public. Responses were generated using the publicly accessible version of each AI system available at the time of data collection.

## Comparator evidence

AI-generated responses were compared with findings from the 2024 NHS Adult Inpatient Survey (1). The primary comparator was the survey's quantitative results. A secondary comparator was the National Qualitative Report, which provides thematic analysis of patients' free-text comments (10).

The NHS Adult Inpatient Survey was selected because it provides a comprehensive and authoritative account of inpatient experience that is widely used by NHS organisations, regulators, and policymakers to monitor quality and inform service improvement.

Other sources of patient experience data (such as the NHS Friends and Family Test (FFT), Patient Reported Outcome Measures (PROMs), and complaints data) exist in England, but the analysis was limited to the NHS Adult Inpatient Survey to ensure a clearly defined scope.

## Prompts

Each AI system was asked the same three questions:

1. What are the experiences of adults receiving NHS acute inpatient care in England in 2024?
2. What does the 2024 NHS Adult Inpatient Survey show about patients' experiences of discharge from hospital?
3. What does the 2024 NHS Adult Inpatient Survey show about differences in patient experience between different groups of patients?

The first prompt captures general representations of inpatient experience. The second and third prompted AI systems to summarise findings from a specific evidence source, focusing on discharge experiences and inequalities respectively.

Data were collected on 26th June 2026. Prompts were entered through the standard user interface of each platform using new accounts and new conversation windows. All AI systems were used with default settings.

The full responses generated by each AI system are provided in a separate Appendix document.

## Analysis

Responses were retained in their original form and analysed using a structured coding framework. The analyses examined:

- **Alignment with evidence**: whether statements were supported by the survey findings.
- **Thematic representation**: which aspects of patient experience and which patient groups were discussed. Coding was conducted by the researchers and drew on the Picker Principles of Person Centred Care (11).
- **Framing**: how findings were presented, including tone and perspective.
- **Omissions and distortions**: aspects of the evidence that were left out, underrepresented, or materially reframed.
- **Relative prominence**: the amount of attention given to different themes within responses.

Consistent with the report's focus on evidence translation, the analysis examined not only factual accuracy but also how AI systems selected, interpreted, and presented patients' reported experiences.

# Findings

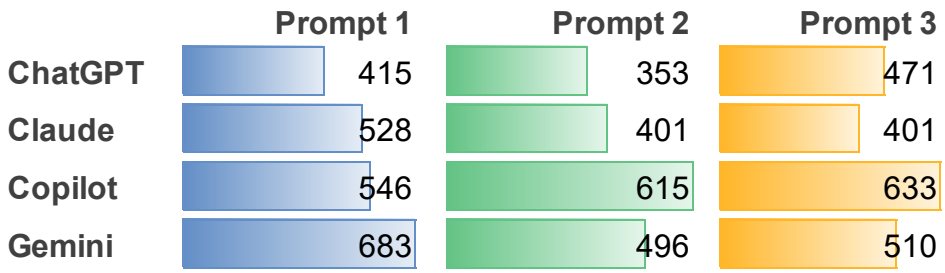
Findings are presented in two parts. First, the report explores patterns that were consistent across AI systems and prompts. It then examines each prompt in turn, focusing on thematic representation, subgroup representation, and alignment with the underlying survey findings.

Several patterns were evident across AI systems and prompts. Differences between AI systems were driven less by the evidence they drew on and more by how they selected, framed, and interpreted that evidence. Common patterns included a preference for quantitative findings over qualitative evidence, differences in the prominence given to particular themes, and variation in how wider explanations and narratives were applied to the survey findings.

## Response length

Response length varied between AI systems. Across the three prompts, ChatGPT generated the shortest responses (mean 413 words), followed by Claude (mean 443 words), while Gemini (mean 563 words) and Copilot (mean 598 words) produced substantially longer responses. However, longer responses did not necessarily provide more context or interpretation. In Copilot's case, additional length reflected formatting and presentation choices.

*Figure 1. Response length (word count) by AI system and prompt*




## Cited sources

The AI systems often used the same sources. Across all prompts, the organisations cited most frequently were the Care Quality Commission (CQC), NHS England, and Picker. This suggests that the differences between responses were not mainly due to using different sources of evidence. Instead, the differences were more likely to come from how each AI system chose to summarise the same information.

*Table 1. Sources cited across AI systems*

| Source | Systems cited |
|---|---|
| CQC | ChatGPT, Claude, Gemini, Copilot |
| NHS England | ChatGPT, Claude, Gemini, Copilot |
| Picker | ChatGPT, Claude, Gemini |
| Other organisations (e.g. Carers UK, Healthwatch, Pharmaceutical Journal, Nuffield Trust) | ChatGPT, Claude, Gemini |

Most of the citations were to summary webpages and published reports rather than the original data tables. This was especially true for the qualitative findings because the original patient comments are not publicly available. In many cases, the AI systems were therefore relying on information that had already been summarised and interpreted before producing their own responses.

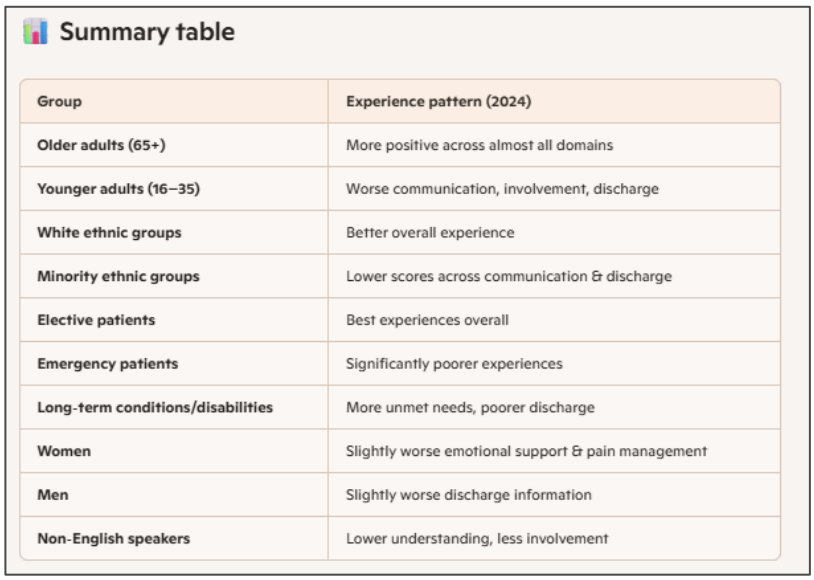
This suggests that AI-generated summaries are often one more step in the process of summarising evidence. By the time information reaches an AI system, it has already been summarised and interpreted in survey reports, webpages and other publications. As a result, differences between AI-generated responses may reflect both the choices made by the AI system and the way the original evidence was presented in the sources it used.

The way the AI systems cited their sources also varied. Most responses included at least one citation, but Microsoft Copilot did not provide any citations for the two follow-up prompts about discharge experiences and differences between patient groups. Across all four AI systems, the number of citations was highest for Prompt 1, lower for Prompt 2, and lowest for the prompt about differences between patient groups (Prompt 3). Without citations, it is much harder to tell which statements come directly from published evidence and which reflect the AI system's own interpretation.

## Presentation of responses

AI systems differed in how they presented information. Claude mainly used narrative text. ChatGPT and Gemini used a mix of narrative paragraphs and bullet points, with Gemini also making use of numbered sections to organise content. In contrast, Copilot made extensive use of headings, bullet points and summary tables. These differences in structure could influence which aspects of patient experience attract users' attention.

*Figure 2. Copilot Prompt 3 response, use of a summary table*



Summary table

| Group | Experience pattern (2024) |
|---|---|
| Older adults (65+) | More positive across almost all domains |
| Younger adults (16–35) | Worse communication, involvement, discharge |
| White ethnic groups | Better overall experience |
| Minority ethnic groups | Lower scores across communication & discharge |
| Elective patients | Best experiences overall |
| Emergency patients | Significantly poorer experiences |
| Long-term conditions/disabilities | More unmet needs, poorer discharge |
| Women | Slightly worse emotional support & pain management |
| Men | Slightly worse discharge information |
| Non-English speakers | Lower understanding, less involvement |

## Use and presentation of evidence

Across all four AI systems, quantitative evidence was generally given more attention than qualitative evidence.

For the first prompt (Prompt 1), all four AI systems presented quantitative evidence before discussing qualitative evidence. This meant that users were first introduced to patient experience through numerical findings.

Gemini relied most heavily on the quantitative findings and made only limited use of the qualitative evidence. ChatGPT and Claude both discussed themes from the qualitative evidence, but these appeared later in their responses. Copilot took a slightly different approach, integrating quantitative and qualitative evidence together throughout its response.

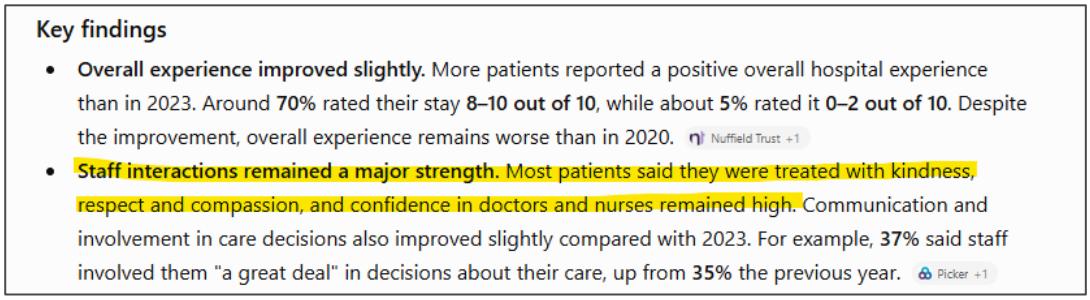
This emphasis on quantitative findings is not surprising. The Adult Inpatient Survey is primarily based on structured questions and statistical reporting, meaning that most of the available evidence is quantitative. It may also suggest AI systems treat survey statistics as stronger evidence than patients' written accounts. Another possible explanation is that quantitative findings are simply easier for AI systems to summarise in a concise response.

For the third prompt (Prompt 3), it was often less clear which evidence was being used. Percentage figures were not reported, making it difficult to tell whether a statement was based on the survey results, patients' written comments, or the AI system's own interpretation. This was made more difficult because ChatGPT and Copilot did not provide links to their sources for Prompt 3, making it harder to check the evidence behind their responses.

There were differences in how AI systems presented quantitative findings. For the first prompt, Gemini and Claude more often included specific percentages alongside their statements, making it easier to see how their conclusions related to the survey results. ChatGPT and Copilot occasionally described the findings without including supporting figures.

For example, ChatGPT referred to patients being treated with kindness, respect and compassion without providing supporting survey results. Similarly, Copilot used phrases such as "*Most felt able to talk to staff about worries or fears*" and "*Some reported issues with food, noise, cleanliness, and call-bell response times*" without saying what proportion of patients had these experiences. Without this context, it is difficult for readers to judge how common these experiences were or how much weight they should be given.

*Figure 3. ChatGPT Prompt 1 response referencing kindness, response and compassion without a linked numerical result*

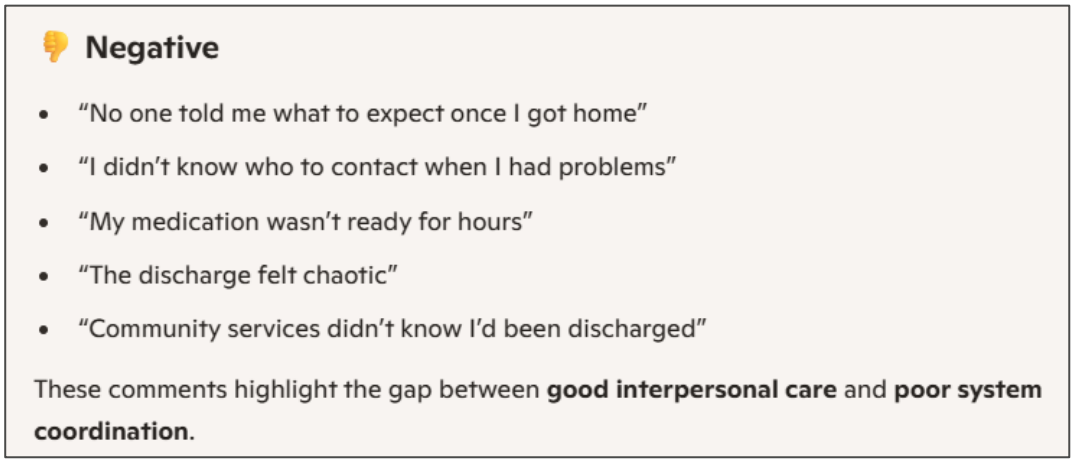

Key findings

- **Overall experience improved slightly.** More patients reported a positive overall hospital experience than in 2023. Around **70%** rated their stay **8–10 out of 10**, while about **5%** rated it **0–2 out of 10**. Despite the improvement, overall experience remains worse than in 2020. Nuffield Trust +1
- **Staff interactions remained a major strength.** Most patients said they were treated with kindness, respect and compassion, and confidence in doctors and nurses remained high. Communication and involvement in care decisions also improved slightly compared with 2023. For example, **37%** said staff involved them "a great deal" in decisions about their care, up from **35%** the previous year. Picker +1

## Representation of patient voice

The AI systems differed in how they presented the qualitative evidence. ChatGPT, Claude and Gemini summarised the qualitative findings by describing the main themes. Copilot took a different approach. In response to Prompt 2, it included what appeared to be direct quotations from patients.

*Figure 4. Copilot Prompt 2 response using patient quotes*


👎 **Negative**

- "No one told me what to expect once I got home"
- "I didn't know who to contact when I had problems"
- "My medication wasn't ready for hours"
- "The discharge felt chaotic"
- "Community services didn't know I'd been discharged"

These comments highlight the gap between **good interpersonal care** and **poor system coordination.**

The quotation marks made it appear that these were patients' own words, making the response feel more authentic. Inclusion of such quotations may have the effect of helping readers connect more directly with patients' experiences. However, when compared with the original source material, these did not appear to be genuine patient quotations. Instead, they seemed to have been generated by the AI system based on themes in the qualitative report. Although the meaning was broadly consistent with the published findings, the wording could not be matched to any patient comments in the source material.

This matters because quotations are often treated as direct evidence of patients' experiences. By generating rather than reproducing quotations, the AI system blurred the distinction between patient testimony and AI interpretation. As a result, qualitative evidence may become further removed from patients' original words and meaning.

## Framing and narrative construction

### Framing of improvement and performance

AI systems sometimes presented the same survey finding in different ways. For example, both Gemini and Copilot reported that 57.9% of patients felt there were always enough nurses on duty. However, Gemini presented this as a sign of improvement, highlighting the increase from 55.7% in 2023 and 51.6% in 2022. In contrast, Copilot included the same finding in a section called "*Areas where patients struggled*" and stated that "*only 57.9%*" of patients felt there were always enough nurses.

Both interpretations were supported by the survey results, but they told different stories. Gemini encouraged readers to focus on improvement over time, while Copilot highlighted the gap between patient expectations and their experience of staffing levels. As a result, the same statistic was used to support two different accounts of inpatient care.

*Figure 5. Gemini Prompt 1 framing of staffing levels*

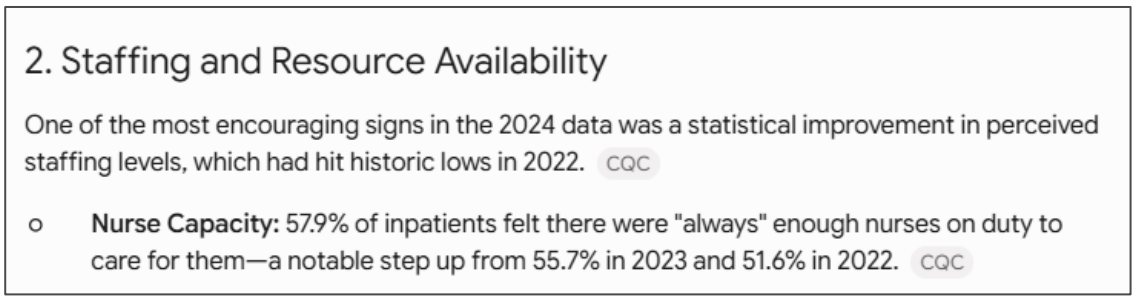
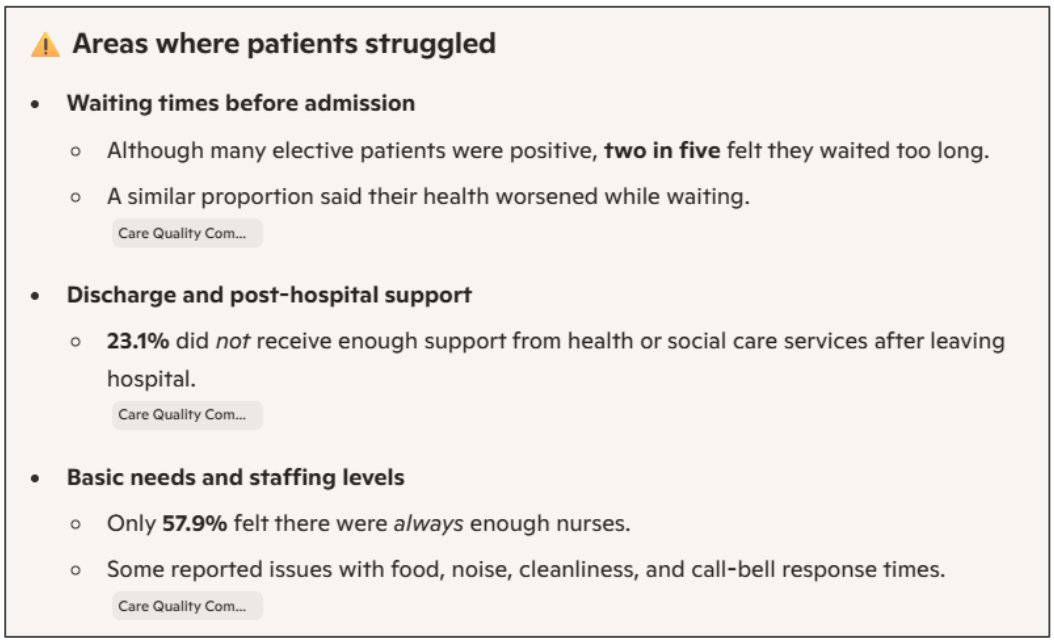


*Figure 6. Copilot Prompt 1 framing of staffing levels*



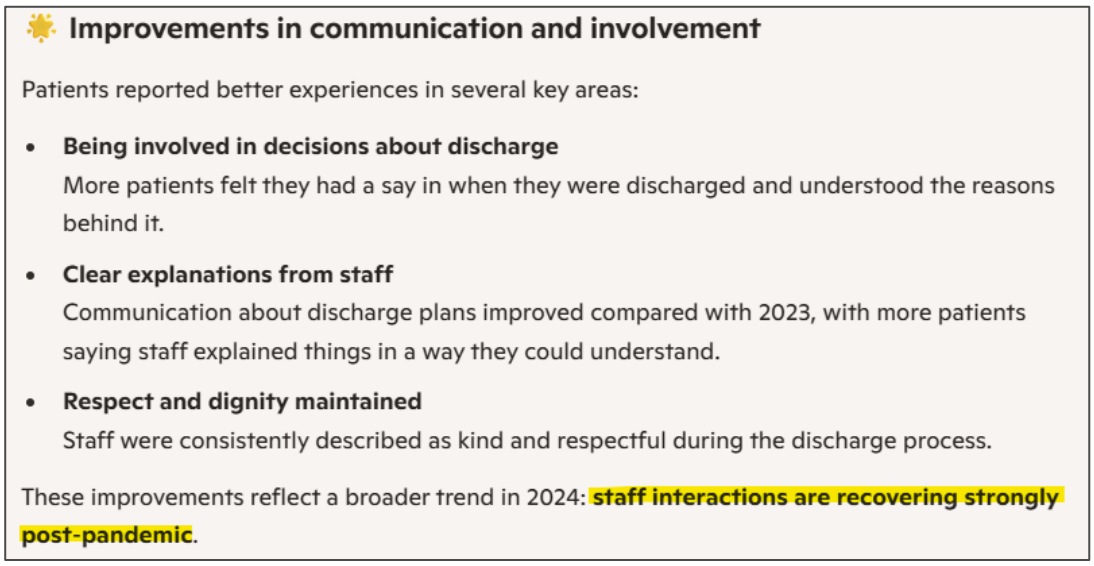

## Use of explanatory narratives

AI systems often placed survey findings within a wider narrative that extended beyond a simple description of results. In many cases, these narratives were not created entirely by the AI systems. Instead, they drew on explanations and framing already present in the source evidence. However, different systems prioritised different contextual explanations and occasionally included connections that were not explicit in the cited sources. As a result, the same survey findings could be framed in different ways across AI systems.

### *COVID-19 pandemic*

One difference between the AI systems was whether they referred to the COVID-19 pandemic. In response to Prompt 1, ChatGPT, Gemini and Copilot all said that patient experience had improved but remained below pre-pandemic levels. Copilot also referred to the pandemic in response to Prompt 2, whereas none of the other AI systems mentioned it in either Prompt 2 or Prompt 3. This framing presented current performance as part of a wider recovery story encouraging users to view ongoing problems in the context of the disruption caused by the pandemic. For example, Gemini stated that *"acute care experiences remain significantly strained compared to pre-pandemic (2020) baselines"*. Claude did not use this framing. Although it compared findings with 2020, it did not explicitly mention the pandemic.

*Figure 7. Copilot Prompt 2 response referencing the pandemic*



**Improvements in communication and involvement**

Patients reported better experiences in several key areas:

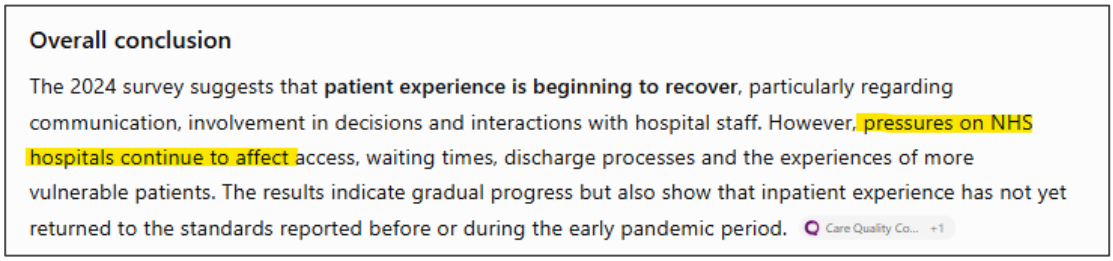
- **Being involved in decisions about discharge**
  More patients felt they had a say in when they were discharged and understood the reasons behind it.
- **Clear explanations from staff**
  Communication about discharge plans improved compared with 2023, with more patients saying staff explained things in a way they could understand.
- **Respect and dignity maintained**
  Staff were consistently described as kind and respectful during the discharge process.

These improvements reflect a broader trend in 2024: **staff interactions are recovering strongly post-pandemic**.

*NHS pressures*

ChatGPT, Gemini, and Copilot all referred to wider pressures on the NHS. ChatGPT said that *"pressures on NHS hospitals"* were affecting access and discharge, while Gemini described *"systemic bottlenecks"* affecting the patient journey. Copilot also identified *"systemic pressures"* as a theme in patients' comments. All three AI systems mentioned these wider pressures in response to Prompt 1, and Copilot also referred to them in response to Prompt 3. These explanations encourage readers to see problems such as waiting times and delays to discharge as part of wider pressures across the health and care system. Claude took a different approach. Although it discussed the same issues it did not link them to wider pressures on the NHS.

*Figure 8. ChatGPT Prompt 1 response referencing pressures on the NHS*



**Overall conclusion**

The 2024 survey suggests that **patient experience is beginning to recover**, particularly regarding communication, involvement in decisions and interactions with hospital staff. However, pressures on NHS hospitals continue to affect access, waiting times, discharge processes and the experiences of more vulnerable patients. The results indicate gradual progress but also show that inpatient experience has not yet returned to the standards reported before or during the early pandemic period. Care Quality Co... +1

*Wider policy context*

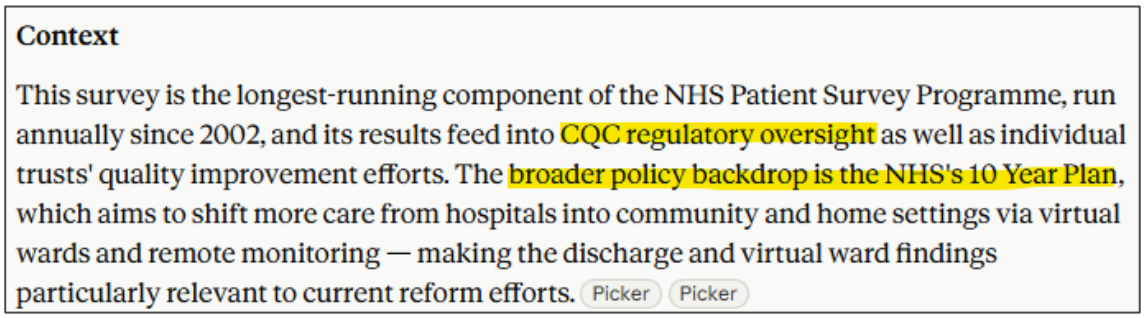
Claude was the only AI system to discuss the survey findings in the context of wider NHS policy. In its responses to both Prompt 1 and Prompt 2, it linked patients' experiences to the NHS Ten Year Plan.

Claude was also the only AI system to discuss how the survey findings are used by regulators. In Prompt 1 it noted that the survey informs CQC regulatory oversight. In

Prompt 3, it linked observed inequalities to the survey's role in identifying risks and informing assessment activity. None of the other AI systems referred to the regulatory use of the survey findings.

By making these links, Claude's responses encourage users to consider not only what patients reported, but also what the findings might mean for NHS policy, regulation and service improvement.

*Figure 9. Claude Prompt 1 response referencing policy context*




An additional observation is that AI systems appeared to select particular explanatory narratives. ChatGPT, Gemini and Copilot tended to emphasise NHS pressures and post-pandemic recovery, whereas Claude highlighted policy and regulatory context. Where source materials contained both perspectives, different AI systems select and emphasised one over the other. This contributed to differences in how the same evidence was presented.

### Framing discharge as a coordination challenge

Across all four AI systems, discharge experiences were framed as an issue of coordination between services in response to Prompt 2 (Table 2). ChatGPT emphasised the need for better coordination between different providers and services. Claude described discharge as part of a broader *"coordination problem"*. Gemini highlighted failures in integration between hospital and social care services, while Copilot referred to poor system coordination.

By describing discharge in this way, issues were presented as symptoms of wider coordination failures between organisations and services. This encouraged readers to see discharge as a systems issue rather than as a collection of separate patient experience issues. While this interpretation was generally consistent with the source material, it reflected an additional layer of synthesis in which AI systems selected and prioritised particular explanations.

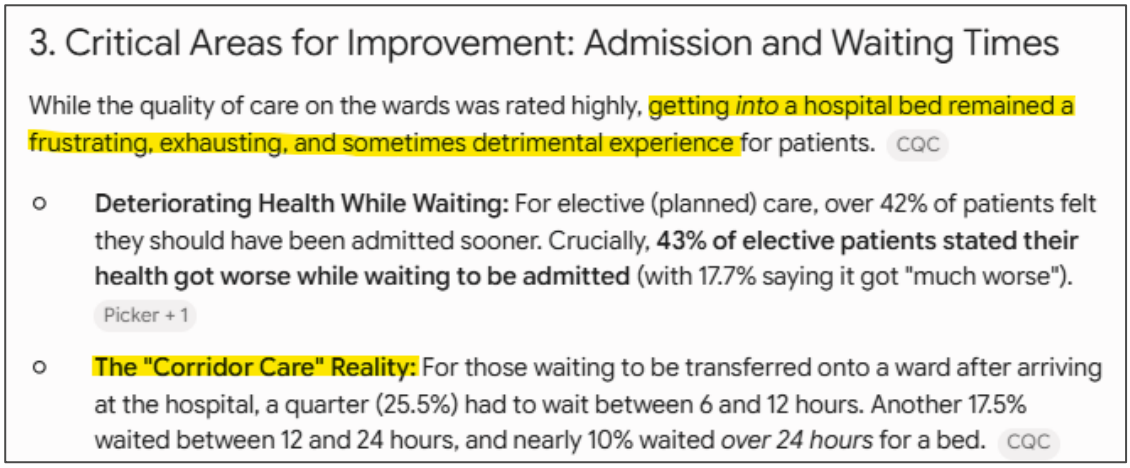
*Table 2. Coordination framing in response to Prompt 2.*

| AI system | Coordination framing |
|---|---|
| ChatGPT | "the survey highlighted the importance of better coordination between hospitals, community services and social care" |
| Claude | "discharge as a symptom of a broader coordination problem." |
| Gemini | "lack of coordinated social care integration" |
| Copilot | "poor system coordination" |

## Language and style

The AI systems used different styles of language to describe findings. Gemini tended to adopt a more expressive, emotive style than the other systems. For example, in Prompt 1 it described admission delays as a *" frustrating, exhausting, and sometimes detrimental experience"*, referred to the *"Corridor Care Reality"*, and characterised the overall findings as *"a tale of two halves"*. In Prompt 2, it used the phrase *"Left Without Help"* when discussing post-discharge support.

*Figure 10. Gemini Prompt 1 response with expressive language*




ChatGPT and Claude generally adopted a more restrained style, using phrases such as *"modest improvements"*, *"concern"*, and *"persistent problem areas"*. Copilot used some descriptive labels but generally presented findings in a more factual way.

These differences in language shaped how survey findings were presented. More emotive or descriptive language can draw attention to particular problems and make them seem more important, while more neutral language may lead readers to see the same findings in a more balanced way.

### Perspective

Across all four AI systems, patients' experiences remained the primary focus. The responses mainly described what patients reported during their hospital stay rather than the views of staff, NHS organisations, or policymakers. Even when the AI systems discussed wider issues, they usually explained them in terms of how they affected patients.

As a result, differences between the AI systems were more evident in how experiences were framed than in whose perspective was prioritised. For example, ChatGPT stated that *"more patients felt they were involved in planning their discharge, but many still reported that arrangements for leaving hospital were not well coordinated, particularly where community or social care support was needed"*. Although this refers to coordination between services, the focus remained on how patients experienced the discharge process.

This is important because patient experience evidence is intended to provide an account of care from the patient's perspective. Despite the differences in how the findings were explained, all four AI systems largely kept that person centred perspective.

Taken together, these findings on narrative construction suggest that the AI systems were doing more than selecting information from the survey: they were also explaining the findings and linking them into broader stories. Similar patterns were seen across several responses, including framing inpatient care as a story of post-pandemic recovery, interpreting discharge through the lens of system coordination, and explaining patient experience in terms of wider NHS pressures. These explanations help readers make sense of the findings, but they also go beyond what was directly reported in the survey.

The following sections examine these patterns in more detail for each prompt.

## Prompt 1 - What are the experiences of adults receiving NHS acute inpatient care in England in 2024?

Although the prompt did not mention the NHS Adult Inpatient Survey, all four AI systems used its findings in their responses. This section looks at how each AI system presented, explained and emphasised those findings.

### Representation of the Picker Principles

The themes identified in the responses were grouped by the researchers using the Picker Principles of Person Centred Care. This made it possible to compare which aspects of patient experience each AI system gave the most attention to. Themes that did not fit within the Picker Principles were also identified.

All four AI systems covered a broadly similar set of Picker Principles in response to Prompt 1. Three additional themes also appeared: staffing, inequalities and overall evaluations of care.

Table 3 shows how much attention each AI system gave to each Principle in response to Prompt 1. Some Principles were not mentioned at all, while others were discussed in detail, making it possible to compare which aspects of inpatient experience each AI system emphasised most.

*Table 3. Elaboration of Principles across AI systems*

| Principle | ChatGPT | Claude | Gemini | Copilot |
|---|---|---|---|---|
| Attention to physical and environmental needs | Absent | Brief | Brief | Moderate |
| Clear information, communication and support for self-care | Moderate | Moderate | Moderate | Moderate |
| Continuity of care and smooth transitions | Moderate | Extended | Extended | Moderate |
| Effective treatment by trusted professionals | Brief | Absent | Absent | Moderate |
| Emotional support, empathy and respect | Moderate | Brief | Moderate | Extended |
| Fast access to reliable healthcare advice | Moderate | Moderate | Extended | Moderate |
| Involvement and support for family and carers | Absent | Brief | Absent | Absent |
| Involvement in decisions and respect for preferences | Moderate | Moderate | Brief | Moderate |
| **Additional themes identified** | **ChatGPT** | **Claude** | **Gemini** | **Copilot** |
| Inequalities | Brief | Absent | Extended | Absent |
| Overall evaluation of care | Moderate | Moderate | Moderate | Moderate |
| Staffing | Brief | Absent | Moderate | Moderate |

Across all four AI systems, the greatest attention was given to overall experience, communication, access to care, and continuity of care. In contrast, family and carer involvement and people's physical and environmental needs received relatively little attention.

Overall experience was usually covered in the opening summary, alongside changes over time. Communication was most often discussed in relation to explanations from staff,

being included in conversations, and information provision at discharge. Access-related themes focused on waiting, admission, and getting help when needed, while continuity of care centred largely on discharge planning and coordination between services.

Gemini devoted the most attention to "Fast access to reliable healthcare advice", which covered issues of waiting times and admission. Claude provided the most detailed discussion of "Continuity of care and smooth transitions". Copilot placed the greatest emphasis on "Emotional support, empathy and respect" whereas ChatGPT gave more balanced coverage across the different Principles.

Some Picker Principles received very little attention. "Involvement and support for family and carers" was mentioned only by Claude in a wider discussion about persistent problem areas. This is likely to reflect the content of the survey itself, which includes less evidence on this topic than on communication, waiting times and discharge.

*Figure 11. Claude Prompt 1 response referencing family and carers*

**Persistent problem areas**

Despite these gains, significant friction points remain at both ends of the hospital journey:

*Getting into hospital:* Two in five people (42%) with a planned admission said they would have liked to be admitted sooner, and 43% said their health got "a bit" or "much worse" while on the waiting list. Picker

*Leaving hospital:* Discharge remains weak. Only half (49%) said they were "definitely" given enough notice about when they would leave, and 40% said staff involved their family or carers in discharge discussions "not very much" or "not at all". Picker

*Virtual wards:* As more care shifts toward hospital-at-home models, 55% of patients said staff "definitely" gave them information about the risks and benefits of continuing treatment on a virtual ward, but 19% said they were not informed of the risks and benefits at all. Picker

Issues relating to attention to physical and environmental needs (such as food, cleanliness, noise, and pain management) only appeared in responses by Claude and Copilot and were generally given much less prominence compared to other themes. This pattern suggests that the AI systems prioritised themes relating to relational aspects of care over more practical aspects of the inpatient experience. However, the same prioritisation was evident in the summary reports and webpages cited by the AI systems. As a result, this emphasis reflects both the way the source material presents the findings and the choices made by AI systems when constructing their responses.

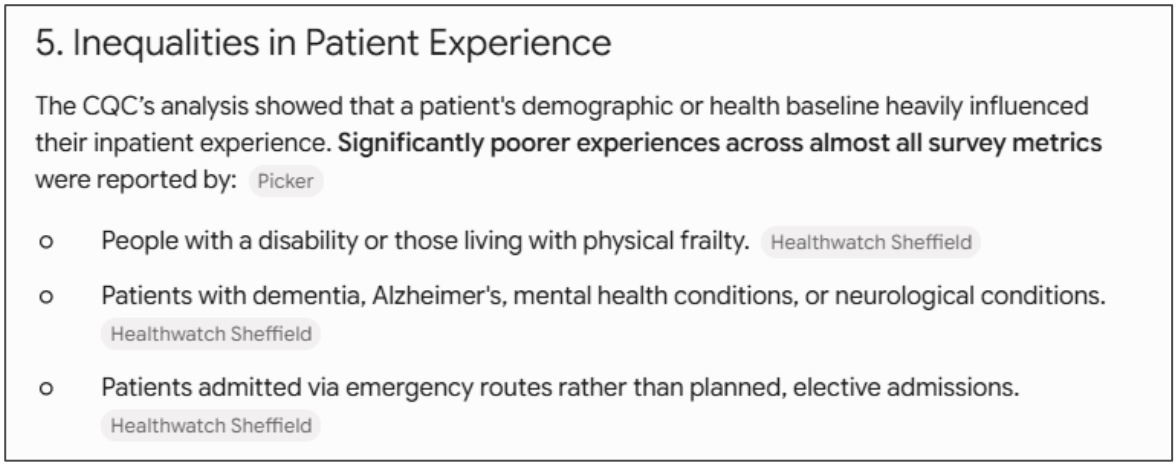
Staffing levels were referenced by all AI systems except Claude in response to Prompt 1. ChatGPT referred to concerns that staffing levels were affecting responsiveness and continuity of care; Gemini highlighted improvements in perceived staffing levels and nurse availability; and Copilot included staffing levels as both a challenge and a factor influencing patients' experiences.

## Patient groups represented

Responses to Prompt 1 largely focused on findings at the national level, describing the average experience of inpatient care across all respondents.

Gemini was the only AI system to include a dedicated section describing how experiences varied for different patient groups. ChatGPT also referred to these differences, but as part of a broader summary of key findings rather than a standalone discussion.

*Figure 12. Gemini Prompt 1 response referencing inequalities*




Claude and Copilot did not discuss variation between patient groups in response to Prompt 1. As a result, users reading these responses might reasonably conclude that the reported strengths and weaknesses of inpatient care applied equally across the patient population. However, the survey findings show that some groups consistently reported poorer experiences than the national average. By explicitly highlighting these differences, Gemini and ChatGPT provided a more complete account of inpatient care and drew attention to inequalities that were largely absent from the other responses.

This is important because summaries focused only on 'average' patient experience can obscure important differences between groups. The omission of these findings may therefore lead users to underestimate the extent of inequalities within inpatient care.

## Prominence of themes

The ordering of themes was broadly consistent across AI systems. All four responses began with an overall assessment of inpatient care before moving on to more specific themes. These opening assessments typically combined evidence of improvement with acknowledgement of continuing challenges. They were then followed by more detailed discussion of positive aspects of care, particularly staff interactions, communication and patient involvement. Challenges such as waiting times, discharge difficulties, and inequalities were generally explored later in the responses.

This sequencing may influence how users interpret the findings. By introducing an overall narrative of improvement or recovery before discussing specific problems, the responses positioned challenges within a broader story of progress.

Table 4 illustrates the first three themes introduced by each AI system in response to Prompt 1, demonstrating the extent to which positive aspects of care were prioritised early in the narrative.

*Table 4. First three themes introduced by each AI system*

| AI System | First theme introduced | Second theme introduced | Third theme introduced |
|---|---|---|---|
| **ChatGPT** | Overall evaluation of care | Emotional support, empathy and respect | Effective treatment by trusted professionals |
| **Claude** | Overall evaluation of care | Emotional support, empathy and respect | Involvement in decisions and respect for preferences |
| **Gemini** | Overall evaluation of care | Emotional support, empathy and respect | Fast access to reliable healthcare advice |
| **Copilot** | Overall evaluation of care | Emotional support, empathy and respect | Involvement in decisions and respect for preferences |

## Alignment with patient experience evidence

Overall, alignment with the quantitative survey findings was strong across all four AI systems. Most statistics and trends reported by the AI systems were consistent with the published survey results, with responses generally reproducing the survey's central story of modest improvement alongside continuing challenges in key areas of the patient journey.

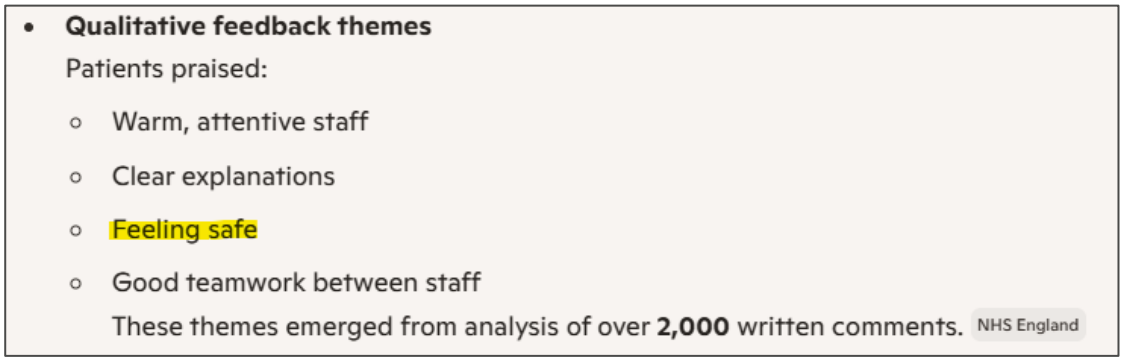
In one case, an AI system repeated an incorrect figure taken from a secondary summary rather than the original survey results.

Alignment with the qualitative evidence was less straightforward than alignment with the quantitative findings. Rather than reproducing findings directly from the qualitative report, the AI systems generally presented thematic summaries. This made it harder to tell which themes came from the published evidence and which reflected the AI system's own interpretation.

For ChatGPT and Claude the relationship between the AI-generated summaries and the underlying qualitative evidence was relatively clear. Copilot's representation was less

straightforward. For example, it appeared to infer broader concepts from the reported themes rather than describing the themes themselves. One example of this is where Copilot identified *"feeling safe"* as a positive aspect of care highlighted by patients.

*Figure 13. Copilot Prompt 1 response referencing "feeling safe"*


- **Qualitative feedback themes**

  Patients praised:
  - Warm, attentive staff
  - Clear explanations
  - Feeling safe
  - Good teamwork between staff

  These themes emerged from analysis of over **2,000** written comments. NHS England

While the qualitative report emphasised themes such as staff kindness, professionalism, and compassion, "*feeling safe*" was not identified as a standalone positive theme. Indeed, references to safety in the qualitative report were associated with areas for improvement, such as patients describing feeling unsafe because of the behaviour of disruptive patients on wards.

Rather than reflecting an explicit finding from the qualitative analysis, Copilot appears to have inferred a sense of safety from other positive staff attributes and care experiences. Such inferences may be reasonable and consistent with the wider findings. However, they blur the distinction between evidence and interpretation, making it harder to judge where reported findings end and AI system generated explanation begins.

Overall, factual alignment with the survey findings was generally strong. However, differences between AI systems were more evident in how findings were interpreted than in the accuracy of the findings themselves.

## Prompt 2 - What does the 2024 NHS Adult Inpatient Survey show about patients' experiences of discharge from hospital?

Unlike Prompt 1, which asked about inpatient experience as a whole, Prompt 2 focused on one stage of the patient journey. As a result, the four AI systems gave more similar responses. All of them focused on communication, discharge planning, and support after leaving hospital. However, they still differed in which aspects of discharge they emphasised and how they explained the issues patients experienced.

### Representation of the Picker Principles

Across all four AI systems, the greatest attention was given to issues of communication, continuity of care, and support after discharge (Table 5). Discharge was commonly represented as a process dependent on coordination between services. Medicines

information received substantial attention from Claude and Gemini, despite being only one component of the wider discharge experience. Family and carer involvement received comparatively less attention than communication, discharge planning and post-discharge support.

Overall, the AI systems showed more agreement in their responses than they did for Prompt 1. This reflects both the narrower focus of the prompt and the concentration of the underlying survey findings on communication, discharge planning, and support after leaving hospital.

*Table 5. Elaboration of Principles across AI systems*

| Principle | ChatGPT | Claude | Gemini | Copilot |
|---|---|---|---|---|
| Attention to physical and environmental needs | Absent | Absent | Brief | Brief |
| Clear information, communication and support for self-care | Extended | Extended | Extended | Extended |
| Continuity of care and smooth transitions | Extended | Extended | Extended | Extended |
| Effective treatment by trusted professionals | Absent | Absent | Absent | Absent |
| Emotional support, empathy and respect | Absent | Absent | Absent | Moderate |
| Fast access to reliable healthcare advice | Absent | Absent | Absent | Absent |
| Involvement and support for family and carers | Brief | Brief | Absent | Absent |
| Involvement in decisions and respect for preferences | Moderate | Brief | Absent | Moderate |
| **Additional themes identified** | **ChatGPT** | **Claude** | **Gemini** | **Copilot** |
| Inequalities | Brief | Absent | Extended | Absent |
| Overall evaluation of care | Brief | Brief | Brief | Absent |

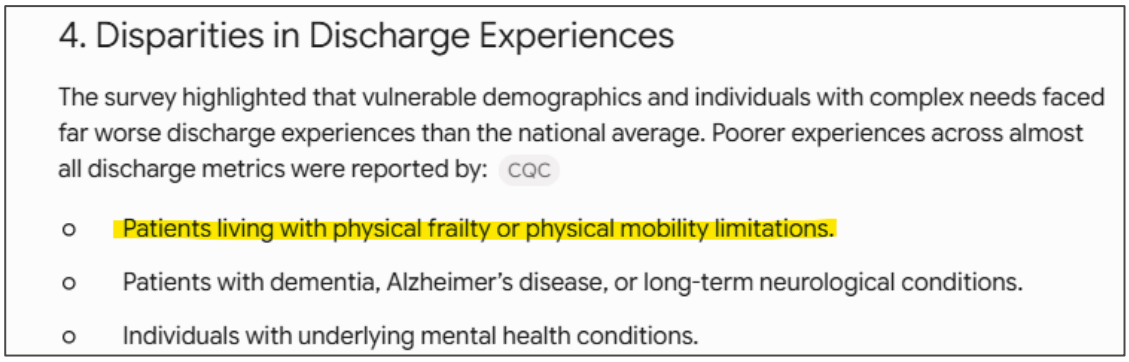
***Note:*** *not all Picker Principles are directly reflected in the survey's discharge measures, and some absences therefore reflect the scope of the underlying evidence rather than omission by the AI systems.*

## Patient groups represented

As with Prompt 1, Gemini was the only AI system to include a dedicated section describing the discharge experiences of different patient groups. ChatGPT made a single reference to the importance of coordination in ensuring that *"older people and those with complex needs"* received appropriate support after leaving hospital. By including a dedicated discussion of variation between patient groups, Gemini highlighted inequalities in discharge experiences that were largely absent from the other responses. However, some of these descriptions extended beyond what could be directly supported by the published survey findings. Claude and Copilot did not discuss variation in discharge experiences between patient groups.

Gemini reported that people living with frailty experienced poorer outcomes across discharge measures, which was consistent with the survey findings. However, some of its broader descriptions of other patient groups were less closely aligned with the underlying evidence. For example, it stated that people with "physical mobility limitations" reported poorer experiences across almost all discharge metrics. While this may have been intended as a reformulation of findings relating to disability, the evidence does not clearly support the extent of the differences described.

*Figure 14. Gemini Prompt 2 response referencing "frailty" and "physical mobility limitations"*


4. Disparities in Discharge Experiences

The survey highlighted that vulnerable demographics and individuals with complex needs faced far worse discharge experiences than the national average. Poorer experiences across almost all discharge metrics were reported by: CQC

- Patients living with physical frailty or physical mobility limitations.
- Patients with dementia, Alzheimer's disease, or long-term neurological conditions.
- Individuals with underlying mental health conditions.

## Prominence of themes

There was less consistency in the ordering of themes for Prompt 2 than was seen in Prompt 1. ChatGPT, Claude and Gemini began with a short overall assessment of discharge experiences before moving to more specific themes, while Copilot moved directly into communication-related issues.

Despite these differences, communication and involvement were consistently prioritised early in the responses (Table 6). ChatGPT and Gemini both introduced communication as the second theme discussed, while Copilot opened with communication and followed this with emotional support and involvement. Claude differed slightly by giving greater prominence to emotional support and family and carer involvement.

*Table 6. First three themes introduced by each AI system*

| AI system | First theme introduced | Second theme introduced | Third theme introduced |
|---|---|---|---|
| **ChatGPT** | Overall evaluation of care | Clear information, communication and support for self-care | Involvement in decisions and respect for preferences |
| **Claude** | Overall evaluation of care | Clear information, communication and support for self-care | Involvement and support for family and carers |
| **Gemini** | Overall evaluation of care | Clear information, communication and support for self-care | Continuity of care and smooth transitions |
| **Copilot** | Clear information, communication and support for self-care | Emotional support, empathy and respect | Continuity of care and smooth transitions |

### Alignment with patient experience evidence

The wording of the prompt is important when interpreting these findings. Unlike Prompt 1, Prompt 2 explicitly asked what the 2024 NHS Adult Inpatient Survey showed about patients' experiences of discharge from hospital. As can be seen in Table 7, all four AI systems therefore positioned their responses as summaries of survey findings.

*Table 7. How AI systems positioned their Prompt 2 responses*

| AI System | Opening framing |
|---|---|
| **ChatGPT** | "The 2024 NHS Adult Inpatient Survey found that patients' experiences of leaving hospital improved slightly compared with 2023, but discharge remains an area where many patients continue to experience problems." |
| **Claude** | "The 2024 survey points to discharge as one of the weakest parts of the patient journey, despite improvements elsewhere in the survey overall." |
| **Gemini** | "The 2024 NHS Adult Inpatient Survey (published by the Care Quality Commission) highlights that the process of leaving the hospital remains one of the most challenging and poorly rated parts of the patient journey." |
| **Copilot** | "The 2024 NHS Adult Inpatient Survey paints a very clear—and often sobering—picture of what adults in England experience when they are discharged from hospital…Below is a structured, detailed breakdown of what the survey shows, grounded in the official 2024 results." |

In this context, readers would reasonably expect statements within the response to be grounded in evidence reported by the survey. However, there were several instances where the AI systems introduced interpretations or additional detail that could not be clearly linked to the underlying evidence.

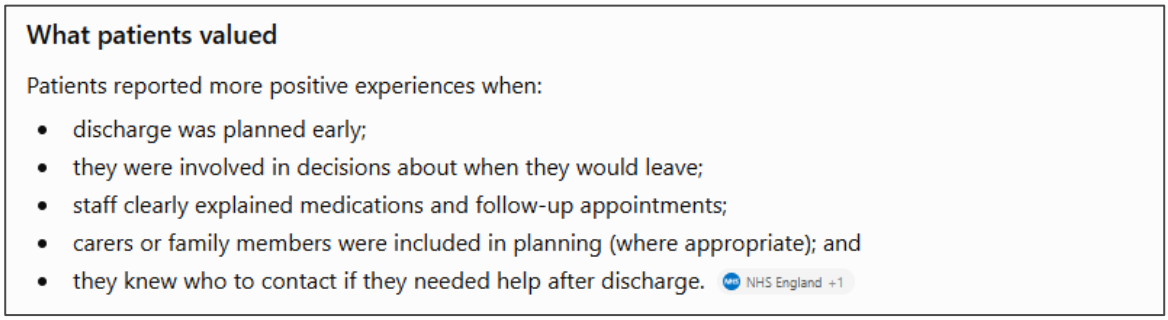
As was seen with Prompt 1, most statistics reported by the AI systems were consistent with the published survey results. However, two AI systems repeated an incorrect figure originating from a secondary summary source rather than the underlying survey findings. This demonstrates how inaccuracies can be reproduced when AI systems draw on already-interpreted evidence sources.

Across both ChatGPT's and Copilot's responses, there were occasions where the distinction between quantitative findings, qualitative evidence, and AI system generated interpretation was not clearly signposted. This ambiguity made it more difficult to assess the evidential basis of some statements. Nevertheless, there were examples where claims could not be clearly aligned with either the quantitative survey results or the accompanying qualitative report. For example, Copilot stated that patients experienced *"difficulty accessing GP appointments or district nursing"* as part of its discussion of persistent problems following discharge. No survey question addressed access to GP appointments or district nursing services, and no equivalent theme was identified in the qualitative report. The evidential basis for this claim was therefore unclear.

Another example is ChatGPT's section on *"What patients valued"*, shown in the image below. While the factors identified are reflected in the survey content, the response implies that these factors were associated with more positive patient experiences, a relationship

that was not explicitly established in the source material. This gives the impression of an additional layer of analysis linking specific aspects of discharge to overall patient experience.

*Figure 15. ChatGPT Prompt 2 response section on "What patients valued"*



**What patients valued**

Patients reported more positive experiences when:

- discharge was planned early;
- they were involved in decisions about when they would leave;
- staff clearly explained medications and follow-up appointments;
- carers or family members were included in planning (where appropriate); and
- they knew who to contact if they needed help after discharge. NHS England +1

As mentioned earlier in this report, Copilot was the only AI system that appeared to make use of direct quotations from patients. This was seen in its response to Prompt 2. Although these statements were broadly consistent with themes in the qualitative report, they were not identifiable patient quotations. As a result, the distinction between patients' own words and AI-generated interpretation was less clear.

## Prompt 3 - What does the 2024 NHS Adult Inpatient Survey show about differences in patient experience between different groups of patients?

Unlike Prompt 1 and Prompt 2, Prompt 3 focused on variation in patient experience between different groups of patients. This resulted in all four AI systems organising their responses around inequalities in care, although there were notable differences in the groups discussed and the level of attention given to each.

### Patient groups represented

As shown in Table 8, all four AI systems identified variation in patient experience between different groups of patients, but they differed in which inequalities they prioritised. Age, sex/gender, disability, long-term conditions and route of admission were discussed consistently across all four responses.

*Table 8. Coverage of patient groups by each AI system*

| Patient group | ChatGPT | Claude | Gemini | Copilot |
|---|---|---|---|---|
| Age | ✓ | ✓ | ✓ | ✓ |
| Autism | ✓ | ✗ | ✗ | ✗ |
| Dementia/Alzheimer's | ✗ | ✓ | ✓ | ✗ |
| Diagnosis (ICD-10 disease chapter groups) | ✗ | ✗ | ✗ | ✗ |
| Disability | ✓ | ✓ | ✓ | ✓ |
| Elective/planned admissions | ✓ | ✓ | ✓ | ✓ |
| Emergency admissions | ✓ | ✓ | ✓ | ✓ |
| Ethnicity | ✓ | ✗ | ✗ | ✓ |
| Frailty | ✓ | ✓ | ✓ | ✗ |
| Index of Multiple Deprivation (IMD) decile | ✗ | ✗ | ✗ | ✗ |
| Language and communication needs | ✗ | ✗ | ✗ | ✓ |
| Learning disability | ✓ | ✗ | ✗ | ✗ |
| Length of stay | ✗ | ✗ | ✗ | ✗ |
| Long-term conditions | ✓ | ✓ | ✓ | ✓ |
| Medical/surgical treatment type | ✗ | ✗ | ✗ | ✗ |
| Mental health conditions | ✓ | ✓ | ✓ | ✗ |
| Neurological conditions | ✗ | ✓ | ✓ | ✗ |
| Physical mobility limitations | ✗ | ✓ | ✓ | ✗ |
| Religion | ✗ | ✗ | ✗ | ✗ |
| Sex/Gender | ✓ | ✓ | ✓ | ✓ |
| Sexual orientation | ✗ | ✗ | ✗ | ✗ |

***Note:*** *some categories shown in the table overlap. They are retained as separate groups because the AI systems used different terminology and levels of specificity when describing variation in patient experience.*

ChatGPT, Claude, and Gemini discussed frailty as a source of variation in patient experience. ChatGPT included a dedicated section on frailty, Claude identified frailty as one of the strongest predictors of poor experience, and Gemini discussed physical frailty

and mobility limitations as a distinct patient group. In contrast, Copilot did not mention frailty.

ChatGPT was the only AI system to discuss the experiences of people with autism and people with a learning disability. None of the other AI systems referred to these findings.

Copilot was the only AI system to include a dedicated discussion of language and communication needs, highlighting the experiences of non-English speakers. This introduced a dimension of inequality that was absent from the other three responses. Whether findings for this group were supported by the underlying survey findings is considered in the following section on evidence alignment.

No AI system represented the full range of subgroup analyses included in the survey with several patient groups not mentioned at all by any of the AI systems. These included deprivation (IMD decile), religion, sexual orientation, length of stay, medical/surgical treatment type and diagnosis groupings.

The AI systems differed in how they organised the findings about different patient groups. ChatGPT, Gemini, and Copilot grouped their responses by patient characteristics, with separate sections for age, sex, disability and admission route etc. This encourages readers to think about inequalities by looking at each patient group in turn.

Claude organised the findings differently. Instead of grouping them by patient characteristics, it first discussed the groups that reported poorer experiences and then the groups that reported better experiences. As a result, its response focused on patterns of disadvantage and advantage rather than on individual patient groups.

### Alignment with patient experience evidence

As with Prompt 2, the wording of the prompt is important when interpreting these findings. Prompt 3 asked specifically what the 2024 NHS Adult Inpatient Survey showed about differences in patient experience between different groups of patients.

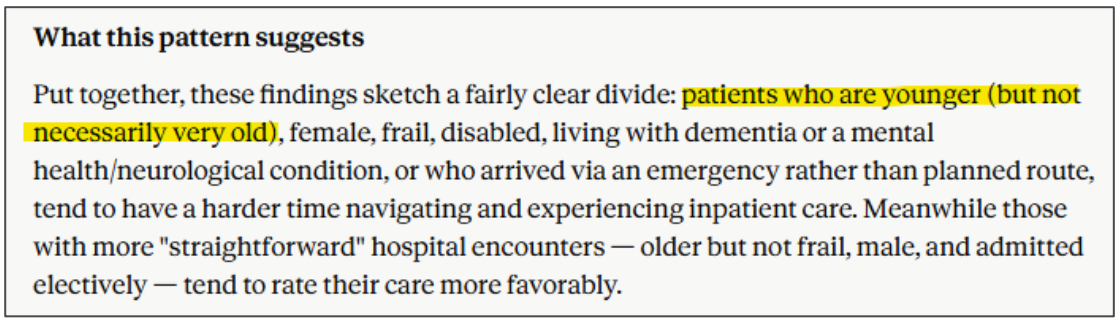
All four AI systems framed their responses as descriptions of survey findings and subgroup analysis. Readers would therefore reasonably expect the statements that follow to be grounded in evidence reported within the survey. However, across all four AI systems, responses to Prompt 3 showed the weakest alignment with the underlying evidence. Compared with Prompt 1 and Prompt 2, the responses contained more generalisations and were more likely to move beyond the specific survey findings, despite the prompt asking directly about evidence from the Adult Inpatient Survey.

The strongest evidence in the Adult Inpatient Survey concerned inequalities associated with age, frailty, disability, and admission route. These were also the patient groups most consistently represented across all four AI systems. These patterns were reproduced reasonably consistently across the AI systems. Older patients were generally identified as reporting more positive experiences, while people living with frailty, people with a disability and those admitted as emergencies were typically described as having poorer

experiences. These findings align closely with the survey evidence and were usually presented accurately.

Claude showed the closest overall alignment with the survey evidence, although elements of its interpretive summaries lacked clarity. For example, it concluded that patients who were *"younger (but not necessarily very old)"* tended to have a harder time experiencing inpatient care. The underlying evidence supports the broader finding that older patients reported more positive experiences than younger patients, but the qualification "*but not necessarily very old*" is difficult to interpret.

*Figure 16. Claude Prompt 3 response referencing younger patients*



**What this pattern suggests**

Put together, these findings sketch a fairly clear divide: patients who are younger (but not necessarily very old), female, frail, disabled, living with dementia or a mental health/neurological condition, or who arrived via an emergency rather than planned route, tend to have a harder time navigating and experiencing inpatient care. Meanwhile those with more "straightforward" hospital encounters — older but not frail, male, and admitted electively — tend to rate their care more favorably.

For the other AI systems, their narratives were only partially supported by the evidence. This includes ChatGPT's treatment of the learning disability subgroup. ChatGPT's response grouped together people with a mental health condition, learning disability, autism, or other long-term conditions, stating that they were more likely to report poorer overall care, poorer communication, weaker discharge planning and unmet needs. The subgroup findings did not support applying this pattern consistently across all of these groups, and the inclusion of learning disability within this broader narrative could not be clearly supported by the subgroup findings reported in the survey.

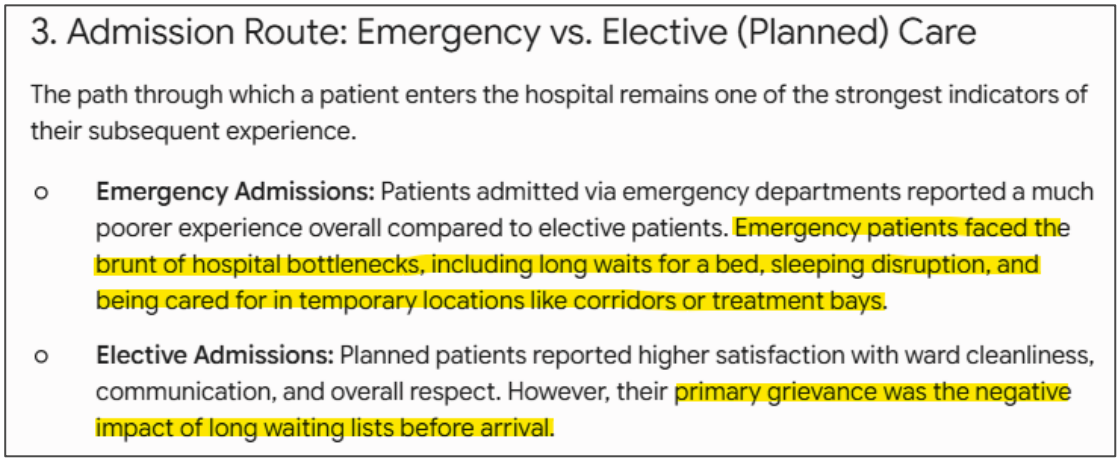
Gemini often translated survey measures into broader concepts that extended beyond those directly captured by the survey. For example, in its discussion of dementia and Alzheimer's disease, Gemini stated that patients reported *"feeling safe"*. The underlying subgroup analysis, however, referred to questions about feeling able to talk about worries and fears, rather than feeling safe. While the concepts are related, the introduction of safety changes the framing of the finding.

Gemini introduced explanatory narratives around type of admission. It suggested that people with an emergency admission experienced the *"brunt"* of hospital bottlenecks. Although this explanation may be plausible, it is not a finding that can be derived directly from the published survey results. Furthermore, the response linked emergency admissions to specific experiences such as sleep disruption and being cared for in temporary locations such as corridors or treatment bays, issues that were not included in the published subgroup analysis.

Gemini also stated that elective patients' *"primary grievance"* was the negative impact of long waiting lists before admission. Although concerns about waiting times were evident in the findings, the survey does not measure the relative importance of different problems to patients and therefore cannot establish whether waiting lists represented their principal concern.

These examples also reflect Gemini's tendency to use more expressive language, identified earlier in the analysis.

*Figure 17. Gemini Prompt 3 response referencing admission route*



3. Admission Route: Emergency vs. Elective (Planned) Care

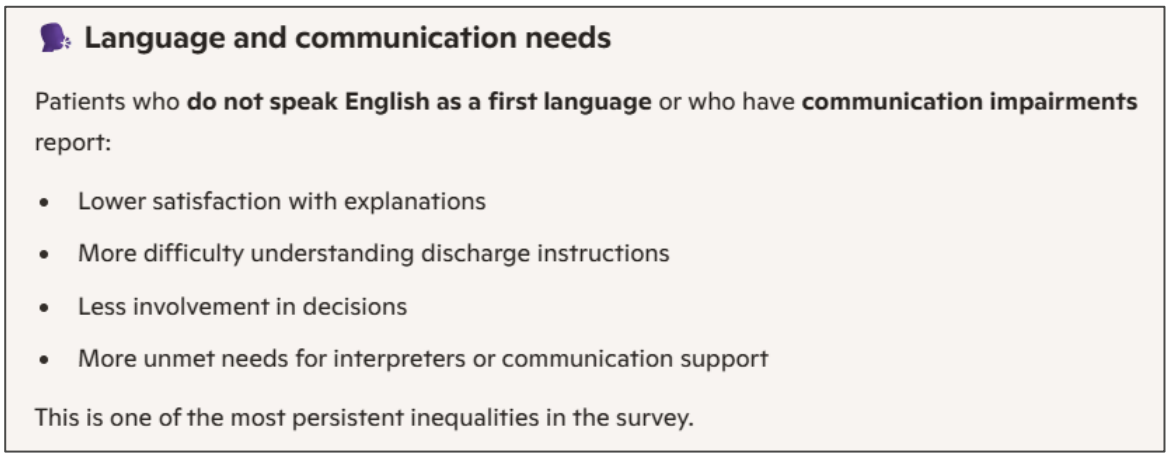
The path through which a patient enters the hospital remains one of the strongest indicators of their subsequent experience.

- **Emergency Admissions:** Patients admitted via emergency departments reported a much poorer experience overall compared to elective patients. Emergency patients faced the brunt of hospital bottlenecks, including long waits for a bed, sleeping disruption, and being cared for in temporary locations like corridors or treatment bays.
- **Elective Admissions:** Planned patients reported higher satisfaction with ward cleanliness, communication, and overall respect. However, their primary grievance was the negative impact of long waiting lists before arrival.

Copilot showed the greatest divergence from the underlying survey evidence. Several sections of the response introduced patterns that could not be clearly linked to the underlying analysis. For example, Copilot stated that emergency patients reported poorer experiences regarding "*information about tests*", despite this not being reported within the subgroup analysis. Similarly, it claimed that men reported poorer experiences in relation to "*understanding medication changes*", even though this was not a question included in the subgroup comparisons. These examples suggest that the AI system moved beyond summarisation and generated additional narrative detail that could not be clearly substantiated from the evidence reviewed.

Copilot also introduced a section on language and communication needs, including claims about "*patients who do not speak English as a first language*" experiencing poorer explanations, less involvement in decisions and greater unmet needs for communication support. These issues were not identified within the evidence reviewed for Prompt 3. Indeed, there was no question in the Adult Inpatient Survey capturing whether respondents spoke English as their first language and no corresponding subgroup analysis reported in the findings. As a result, this section appears to draw on broader narratives about healthcare inequalities rather than findings that could be directly identified within the survey evidence reviewed.

*Figure 18. Copilot Prompt 3 response referencing language needs*




Overall, responses to Prompt 3 showed weaker alignment with the underlying evidence than was observed for Prompt 1 or Prompt 2. While the major patterns of inequality were generally reproduced, AI systems were more likely to introduce explanations, broader subgroup descriptions and additional detail that could not be clearly linked to the published survey findings. This may partly reflect the nature of the underlying subgroup analysis, which is largely presented in tables and charts rather than as a narrative summary. Compared with the findings for overall experience and discharge, this may require AI systems to make greater judgements about what to include and how to explain it.

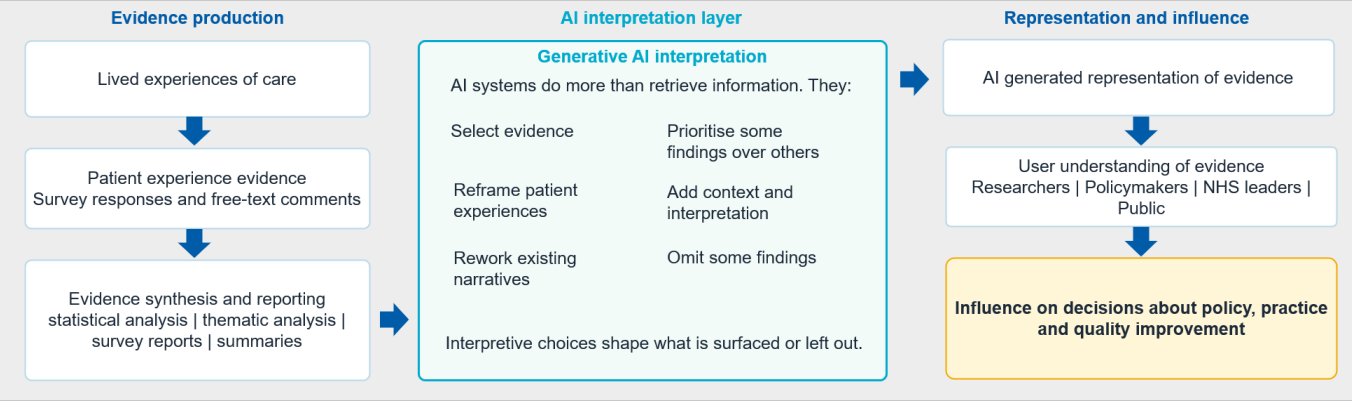
# Conclusion

This report examined how four generative AI systems represented evidence about the experiences of adults receiving NHS acute inpatient care in England, using findings from the 2024 NHS Adult Inpatient Survey as a comparator. It explored how closely AI-generated responses reflected the underlying evidence, which aspects of patient experience were emphasised or left out, and how findings were interpreted and framed.

## AI as an interpreter of patient experience evidence

The findings suggest that generative AI does more than retrieve or summarise patient experience evidence. Instead, it acts as an intermediary in a wider process of evidence translation, shaping how evidence is organised, interpreted, and understood. Figure 19, below, presents a model illustrating this process.

*Figure 19. Process of evidence translation*




Patient experience evidence rarely moves directly from patients to AI systems. Instead, it passes through stages of collection, analysis, and reporting before becoming available to AI systems. Survey responses are analysed, summarised, and communicated through published reports and webpages, which AI systems frequently rely upon rather than primary datasets and data tables. Generative AI then applies a further layer of selection and interpretation as it transforms this evidence into responses. In doing so, it draws not only on patients' reported experiences but also on the ways those experiences have already been framed and interpreted within reports, webpages and other source materials.

This analysis highlights that AI-generated responses should not be viewed as direct representations of patients' experiences. Rather, they represent evidence that has already been analysed and interpreted, with AI contributing an additional stage of translation before that evidence reaches the user. Each stage in this process creates opportunities for

emphasis, omission and reframing, helping to explain why different AI systems can produce different accounts despite drawing on many of the same evidence sources.

## Fidelity to the source evidence varied according to the type of finding being represented

The degree to which findings were preserved varied according to the type of evidence being represented. Headline quantitative survey findings were generally reproduced accurately, particularly where they related to overall experience, major trends, and the most prominent areas of concern.

In contrast, qualitative findings and subgroup analyses were more susceptible to interpretation. As the evidence became more narrative in nature or less readily interpretable, AI-generated responses were more likely to introduce generalisations.

This indicates that the process of evidence translation does not affect all forms of patient experience evidence equally. Some findings remained relatively close to the source material, while others were more vulnerable to reframing.

## The changing nature of patient voice

Patient experience evidence has a unique position within healthcare evidence because its purpose is to communicate how care was experienced by patients themselves. The findings suggest that as this evidence passes through stages of interpretation the patient's original voice becomes progressively abstracted.

This distinction is particularly relevant for qualitative evidence. Patients' comments are first analysed and organised into themes before being summarised in reports and, in some cases, translated again by AI systems into narratives. While each stage may faithfully reflect the overall meaning of the evidence, the resulting account becomes increasingly removed from patients' original words.

This process was most visible where AI systems summarised qualitative findings or generated apparent patient quotations. Although these outputs broadly reflected the themes identified in the evidence, they did not preserve patients' original wording. As a result, readers may encounter responses that feel authentic and patient-led while being several stages removed from the original patient testimony.

## Variation between patient groups was particularly vulnerable to distortion

Responses about the experiences of different patient groups showed weaker alignment with the underlying evidence than responses describing overall patient experience or discharge. This may partly reflect the nature of the evidence itself. Whereas the overall

survey findings and discharge results are presented around relatively clear themes such as communication, waiting times, and discharge planning, subgroup analyses involve multiple comparisons across patient groups and survey measures, often presented in tables and charts rather than as a coherent narrative.

In translating this evidence into conversational responses, AI systems appeared to construct their own organising narratives, selecting which groups to highlight and how their experiences should be characterised. This may help explain why inequalities were more susceptible to generalisation, interpretation and omission than other aspects of patient experience evidence.

## Some aspects of patient experience were consistently amplified while others were diminished

The findings show that AI systems did not give equal attention to all aspects of inpatient experience. Across AI systems, communication, discharge, access to care, and overall evaluations of care were consistently foregrounded. In contrast, physical and environmental aspects of care, including food, hydration, sleep, personal care, privacy, and pain management received comparatively little attention. This pattern is important because representation is shaped not only by what is included, but also by what is left out or given less prominence.

In some cases, these patterns reflected the characteristics of the source material itself. The reports and webpages used by AI systems often placed greater emphasis on these themes and the findings suggest that AI systems frequently reinforced these patterns, producing accounts in which some experiences were consistently amplified while others were marginalised or left out altogether.

As a result, users relying on AI-generated summaries may develop a different understanding of what matters most to patients than they would from engaging directly with the underlying evidence.

## Person centred perspectives were largely preserved

Despite important differences, all four AI systems continued to describe inpatient care primarily through patients' experiences. Even where broader narratives about NHS performance, staffing pressures, policy priorities or service coordination were introduced, these were generally discussed in terms of their implications for patients rather than from the perspective of organisations or healthcare professionals.

This finding is significant because patient experience evidence is intended to provide an account of care from the patient's perspective. While AI systems often reshaped how experiences were presented, they generally retained the person centred orientation of the

underlying evidence. In this respect, the process of evidence translation appeared to alter emphasis and interpretation more than perspective.

## Implications for using generative AI to understand patient experience

Generative AI has the potential to make patient experience evidence more accessible. However, AI-generated answers do not simply present the evidence. They also involve interpretation. This interpretation may come from the AI itself, as well as from the way the original evidence has already been written, summarised, or presented. As a result, AI tools may produce varying answers from the same evidence, highlighting different findings or telling different stories about what the evidence means.

A practical implication is that users should look beyond the AI-generated answer and examine the sources on which it is based. This can help them understand whether the response is drawing on primary evidence or on secondary sources that have already interpreted the evidence for a particular purpose or audience. Where decisions depend on a detailed understanding of patient experience, users should consult the original evidence alongside AI-generated summaries.

Implications also vary depending on how the information is being used.

- For organisations responsible for collecting and reporting patient experience evidence, AI systems are an additional audience. This increases the importance of publishing clear, balanced, and accessible summaries of the evidence.
- For healthcare leaders, patient experience teams, policymakers, and regulators, differences in emphasis and interpretation by AI systems may influence how opportunities for improvement are identified, understood, and prioritised.
- For researchers and analysts, AI-generated accounts may not always reflect the balance or level of detail contained within the underlying evidence.
- For patients and members of the public, different AI systems may present different pictures of the quality of care provided by a hospital or service, even when drawing on the same evidence.

Generative AI can be a valuable tool for accessing and engaging with patient experience evidence, but AI-generated summaries should not be treated as equivalent to the underlying evidence on which they are based. As AI becomes an increasingly common intermediary between evidence and decision makers, understanding how this process shapes perceptions of healthcare quality will become increasingly important.

# Limitations

This study has a number of limitations. First, it provides a snapshot of AI-generated responses at a single point in time, using four of the most popular platforms. Outputs may differ with other platforms and change over time as AI systems, data sources, and settings evolve. This is an important consideration given the rapid pace of change in publicly available AI models: however, we have focused on the broader implications of how a range of AI systems organise and present patient experience evidence, rather than the detailed performance of specific models or platforms.

Second, the analysis focused on three specific prompts, so findings may not generalise to other questions or contexts. Systems can be highly sensitive to variations in prompting: whilst a thorough exploration of this is beyond the scope of this study, we have sought to use simple, plain language prompts that should reasonably reflect common usage and that provide consistency across our testing.

Third, the NHS Adult Inpatient Survey was used as the reference source; other forms of patient feedback may be represented differently.

Fourth, assessing alignment involved judgement about how evidence was selected, interpreted, and presented, and different researchers may have reached different conclusions in some cases.

Finally, responses were generated using new accounts, new conversation windows and default settings. As AI systems become increasingly personalised, real-world responses may vary between users as well as between systems. For example, systems may recall their previous interactions with users and tailor the way data is presented accordingly. This has the potential to add further variation in how evidence is reported back to users, particularly if outputs embed assumptions about the elements of experience that may be of the greatest interest to a specific user. Further testing is required to understand the wider implications of this.

# References


1. Care Quality Commission. NHS Patient Surveys [Internet]. [cited 2025 Sep 11]. Adult Inpatient Survey 2024. Available from: https://nhssurveys.org/surveys/survey/

2. Rabbani SA, El-Tanani M, Sharma S, Rabbani SS, El-Tanani Y, Kumar R, et al. Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. BioMedInformatics. 2025 Sep;5(3):37. doi:10.3390/biomedinformatics5030037

3. Darzi A. High quality care for all: NHS next stage review final report. Lond Dep Health. 2008.

4. Jamieson Gilmore K, Corazza I, Coletta L, Allin S. The uses of Patient Reported Experience Measures in health systems: A systematic narrative review. Health Policy. 2023 Feb 1;128:1–10. doi:10.1016/j.healthpol.2022.07.008

5. Peters U, Chin-Yee B. Generalization Bias in Large Language Model Summarization of Scientific Research [Internet]. arXiv; 2025 [cited 2026 Aug 17]. Available from: http://arxiv.org/abs/2504.00025 doi:10.48550/arXiv.2504.00025

6. Liu F, Zhou H, Gu B, Zou X, Huang J, Wu J, et al. Application of large language models in medicine. Nat Rev Bioeng. 2025 Jun;3(6):445–64. doi:10.1038/s44222-025-00279-5

7. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. 2025 Feb 26;24(1):57. doi:10.1186/s12939-025-02419-0

8. GOV.UK [Internet]. [cited 2026 Aug 19]. AI Skills for Life and Work: General Public Survey Findings. Available from: https://www.gov.uk/government/publications/ai-skills-for-life-and-work-general-public-survey-findings/ai-skills-for-life-and-work-general-public-survey-findings

9. Sollof J. NHSE to roll out Microsoft AI assistant to 505,000 NHS staff. Digital Health [Internet]. 2026 Jun 8 [cited 2026 Jun 23]. Available from: https://www.digitalhealth.net/2026/06/nhse-to-roll-out-microsoft-ai-assistant-to-505000-nhs-staff/

10. NHS England. NHS England [Internet]. [cited 2026 Aug 17]. Adult inpatient experience survey 2024: national qualitative report. Available from: https://www.england.nhs.uk/long-read/adult-inpatient-experience-survey-2024-national-qualitative-report/

11. Picker Principles of Person Centred Care | Picker [Internet]. [cited 2026 Aug 19]. Available from: https://www.picker.org/about-picker/the-picker-principles-person-centred-care

Picker Institute Europe
Suite 6, Fountain House,
1200 Parkway Court,
John Smith Drive,
Oxford OX4 2JY

Tel: +44 (0) 1865 208100

info@pickereurope.ac.uk
picker.org

Charity registered in England and Wales: 1081688
Charity registered in Scotland: SC045048
Company limited by guarantee registered in England and Wales: 3908160