wholesum.

Free-text failure modes: Why human and AI analysts struggle to extract valuable signals from narrative data

A guide to the reasons analysts reach different conclusions, and why it matters for your organisation, based on dozens of real-life examples

Background

From field notes and survey responses to news articles and scientific papers, narrative text often contains rich signals about underlying drivers and outcomes organisations care about. To operationalise such data for understanding and predicting outcomes, teams must first interpret it. Yet interpretations can vary substantially depending on who – or what – is analysing it.

When two human (or AI) analysts read the same piece of text and reach different conclusions, this is often treated as an error to be corrected. The assumption is that with clearer guidance, a better AI model, or more training data, the 'correct' answer can be obtained. However, in practice, interpretive disagreement arises for several distinct reasons. Understanding these causes is crucial if organisations want to avoid unstable and inconsistent analysis.

Across dozens of real-world applications, we have identified 27 recurring failure modes that lead human and AI analysts to interpret the same narrative data differently. These fall into the six broad categories below. Although some may seem straightforward to handle once identified, the real challenge is identifying which ones are present in a given dataset so they can be handled appropriately in downstream analysis.

Six core categories for why interpretation fails

- **A: Information is missing from the text itself (i.e. intrinsic ambiguity)**: the text fundamentally lacks enough information to support an interpretation.
- **B: Information is in the text but gets overlooked**: the necessary detail is present, but is missed or miscounted by the analyst.
- **C: Interpretation depends on patterns within the wider dataset**: the correct meaning is recoverable from wider patterns in the dataset, not one text segment alone.
- **D: Interpretation depends on external context**: the correct reading depends on specialist knowledge or conventions defined externally.
- **E: The text has multiple valid interpretations**: Multiple overlapping causes or poorly defined labels make a definitive interpretation difficult, but can make a probability-based estimate based on prior view and distribution of statements in the data.
- **F: Interpretation depends on a subjective standpoint**: the facts and definitions are known, but the conclusion still rests on a standpoint that reasonable analysts may disagree on.

A. Information is missing from the text itself

In this situation, the text fundamentally lacks enough information to support an interpretation. Hence any interpretation is necessarily subjective, and the only way to get an objective interpretation is to return to the original person who wrote the text, which is often not feasible.

Example: Inherent ambiguity. Text is missing detail, or has genuine ambiguity in writing. For instance, a commercial field note reads "Customer issue." If no other data points refer to an issue, it is not possible to resolve what this means.

Business impact: Analysts may either incorrectly interpret text based on their own prior biases, or incorrectly conclude text is fundamentally unusuable when it reality it can be interpreted when combined with other data points or external knowledge (see B, C, D and E below). Typos and data loss can be particularly common if data is transcribed from speech or extracted from documents.

B. Information is present but gets overlooked

The detail needed is present in the text, but the analyst fails to use it correctly — through inattention, biased weighting, or misreading how a clause or number applies. Given the same text, a careful reader would reach the same conclusion, so disagreement here is a quality problem, not a matter of interpretation.

Example: Positional or framing bias. Information is weighted by where it falls, not by what it says. Early content gets over-weighted, or a late finding gets under-weighted. For instance, an article contains the relevant data at the end of the text, but the analyst does not read that far. Or an outlier data point suggesting no threat leads an analyst to downgrade subsequent observations.

Business impact: Opportunities and threats may be missed or misestimated based on the capability of a specific human analyst or AI model and/or the order they see the data. Datasets with complex multi-entity structure, such as transcripts, comment threads or correlated news reports can further hinder interpretation.

C. Interpretation depends on patterns within the wider dataset

The correct meaning is recoverable from wider patterns in the dataset, not one text segment alone. Interpretation is typically objective if dataset is large and patterns are stable, but it can break down if signals are weak or sparse.

Example: Cross-data linkage. Shared underlying meaning is only interpretable when several data points are seen together. For instance, notes with variants of "MDVs out of stock", "MDCs not in stock" and "Multi-dose vial stockout" can be linked to the same issue, but only if seen together to resolve acronyms and typos.

Business impact: If data points are distributed across a large number of human analysts or AI model runs for analysis, crucial shared context that resolves ambiguities may be missed, with analysts instead reverting to prior biases or assumptions.

D. Interpretation depends on external context

The correct reading depends on information outside the dataset. Interpretation is typically objective once external definitions are known, but it can become subjective if definitions are vague or inconsistently applied.

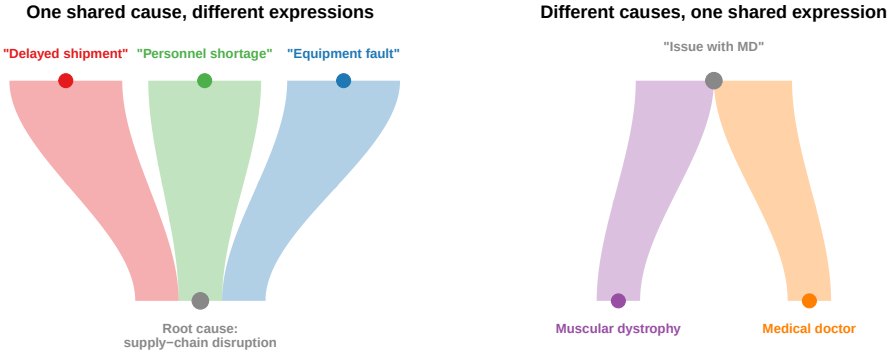
Example: Specialist or internal conventions. The meaning is unambiguous to anyone who knows the convention (e.g. professional jargon or an internal procedure code) but opaque to anyone who doesn't. For instance, "'CR achieved after cycle 2" in a note is clear to oncology staff who recognise it as Complete Response, but meaningless to analysts who do not.

Business impact: Inconsistency in external context may lead to data being incorrectly deemed ambiguous, and/or errors being attributed to inherent limitations of human analysts or AI models. This can lead to expensive (re)classfication costs when the problem is fundamentally a clarity rather than capability issue.

E. The text has multiple valid interpretations.

Multiple overlapping causes or poorly defined labels make a definitive interpretation difficult, and analysts have to make an estimate based on prior view and distribution of statements in the data. There is not a single objective interpretation at the level of the individual response, but the distribution of meaning can often be estimated across the dataset as a whole.

Competing causes. One cause can show up very differently across the dataset, or different real-world causes can produce near-identical For instance, a new supply-chain disruption shows up in three reports as "delayed shipment," "personnel shortage," and "equipment fault." These cannot be definitively linked to the same cause, but it would be misleading to treat as fully independent. Or a note reads "Issue with MD" in a dataset that also mentions problems with muscular dystrophy and medical doctors.



Business impact: If the shared causal structure of a dataset (e.g. field notes from same market, survey from same organisation) is ignored, the number of issues being faced can be dramatically over- or under-estimated, depending on the prior biases of the analyst. Imposing single categories is common when it leads to a routing decision (e.g. in customer service or healthcare triage), but this can further conceal complex, overlapping issues.

F. Interpretation depends on a subjective standpoint

In some situations, the statements might be clear and the definitions known, but the specific conclusion still rests on a normative choice. Unlike D, more information does not necessarily resolve the disagreement, because the disagreement is about values rather than just evidence.

Example: Direction of finding is subjective. Whether something reads as positive or negative depends on standpoint. Example: "Employee vacation days declined sharply this quarter." One analyst reads this as a sign of more focus on work (i.e. positive news), while another reads it as employees being overworked (i.e. negative news).

How Wholesum is tackling these challenges

Wholesum helps organisations make defensible judgements from fragmented evidence. Traditional classification tools are fine-tuned to a narrow, static range of topics or sentiment. Modern large language models are more flexible, but their performance is inconsistent and difficult to operationalise reliably across large, evolving datasets.

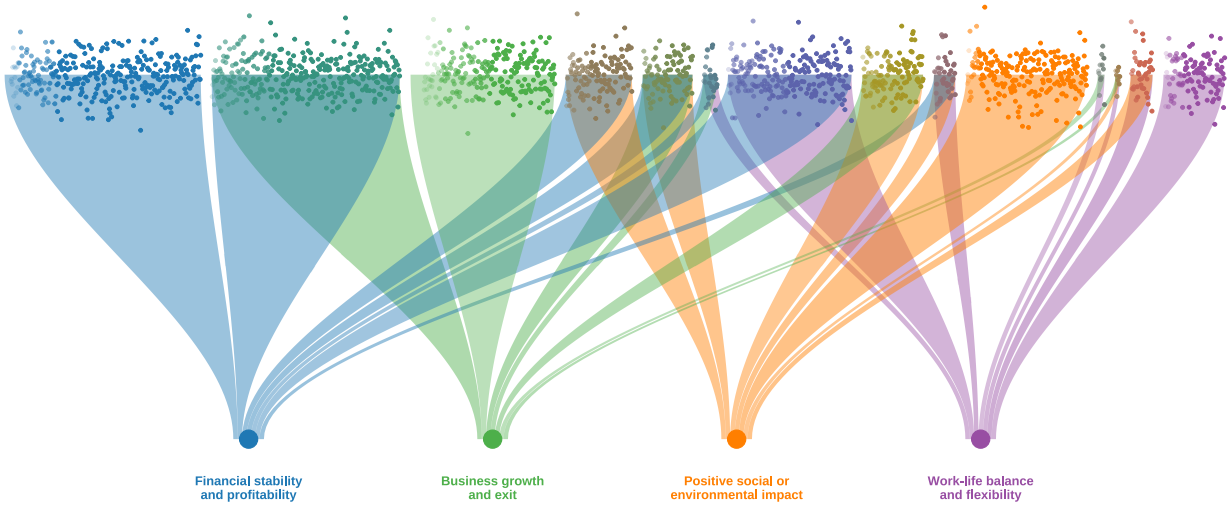
Wholesum takes a different approach, applying comparative statistical inference across an entire dataset, rather than judging each entry in isolation. Rather than treating disagreement as noise to be resolved into a single 'correct' label, this approach identifies which of 27 failure modes within the above categories is likely to be present. We help teams in pharma, finance, national security, adtech, and people insights make that distinction reliably, at scale.

Our work rests on four capabilities: 1) **Full contextual analysis** analyses every entry comparatively, in the context of the wider dataset, surfacing patterns that are undetectable in entry-by-entry analysis; 2) **Failure-mode detection** discovers both widespread and niche disagreement patterns, and quantifies how strong, how common, and how resolvable each one is; 3) **Reproducible outputs** stay consistent across analysis runs and time periods, and are traceable to their source data; 4) **Built for scale**, from thousands to millions of entries, without losing that stability. The result is a deeper understanding that's stable, repeatable, and defensible for high-stakes decisions.

Case study

Wholesum worked with Female Founders Rise, Barclays and Nottingham Business School on a large survey of UK female entrepreneurship. These founders share overlapping markets and funding conditions, so individual survey responses reflected shared causal structure. When asked for their version of long-term success, iterative analysis of the responses converged on four broad themes that were sufficiently specific to distinguish key groups while also broad enough to cover the majority of the survey population.

Because the survey involved short typed responses from an engaged audience answering specific questions, diagnostics indicated low inherent ambiguity (A) and limited need for complex interpretation (B). Instead, the main interpretative challenges were: UK-specific terminology and phrasing that required context from wider dataset (C); reducing ambiguity with relevant external context, particularly on questions about UK-based funding schemes (D); and responses that supported multiple competing explanations (E). Finally, the wording of the theme names themselves reflect an interpretive choice made by the model and subsequently reviewed by humans (F), even if the underlying distinct concepts were stable across runs. The results therefore included uncertainty stemming from the multiple overlapping shared causes within the group, rather than representing a single ground truth.



Each dot represents one founder, grouped by the combination of success themes identified in their response. Dot opacity shows the stability of that theme assignment across the analysis, with lower stability mainly reflecting competing interpretations (E).