

All Hell Broke Loose: A Computational Search for Frustration Associated with Loose, Lost, Stuck, and Floating Equipment in NASA Oral Histories

Jalaj Gangwar

October 2nd 2026

Abstract We report a deliberately narrow computational investigation of a problem that becomes considerably less funny when the object is floating toward a ventilation intake: equipment that is loose, stuck, misplaced, floating, or otherwise unwilling to remain where a crew member put it. The study searched 729 NASA oral-history documents for co-occurrences between equipment-loss/mobility language and frustration or profanity language. The initial lexical screen produced 72 candidate passages. All 72 were re-reviewed by a large language model (LLM), which classified 14 (19.4%) as genuine links between an equipment problem and an expression of frustration; these 14 hits come from 12 documents and describe at most 12 distinct episodes. The retained passages received a mean comedic-value score of 3.36/5, not the 2.1/5 in the project summary, which on recomputation is the mean over all 72 candidates. The most common equipment-side screening terms were “loose” (25), “stuck” (18), and “lost a” (8), while “hell” (28), “damn” (17), and “shoot” (15) dominated the frustration vocabulary. One particularly pleasing lexical accident involved a micrometeorite shield that “came loose” in the same breath as “all hell broke loose.” The study does not estimate how often NASA personnel swear, nor does it establish causation. It instead provides a reproducible way to locate small, narratively rich instances in which physical loss of control and linguistic loss of patience occur in approximately the same neighborhood.

1 Introduction

Spaceflight is unusually intolerant of misplaced objects. They cannot be replaced without a fat check to a Launch Service Provider (LSP). In ordinary terrestrial environments, a screwdriver can be set down, a package can be left on a table, and a loose component can generally be retrieved without changing the coordinate system of the problem. In microgravity, an object can become simultaneously difficult to locate, difficult to restrain, and surprisingly consequential.

This paper examines a small linguistic consequence of that physical problem: whether accounts of loose, stuck, lost, floating, or poorly stowed equipment contain nearby expressions of frustration or profanity.

The motivating observation is not that NASA personnel used profanity. That fact, in isolation, would be neither surprising nor especially measurable; even rocket scientists occasionally discover that “damn” is a perfectly serviceable technical term. Rather, the question is whether a

computationally searchable relation can be found between two otherwise heterogeneous vocabularies: (i) words describing equipment that has escaped intended control, and (ii) words or phrases expressing frustration.

The distinction matters. A passage containing both “loose” and “damn” is not necessarily about a loose piece of equipment causing frustration. “Loose” may describe a schedule, a gate, or an abstract constraint; “damn” may modify an unrelated noun. A computer that sees these two words and immediately declares victory is therefore behaving less like a scientist and more like an overexcited intern with a highlighter. Consequently, lexical co-occurrence is treated here as a candidate-generation mechanism, not as evidence of a genuine causal relationship.

The project therefore used a deliberately over-cautious design. First, a broad keyword screen identified candidate passages. Second, an LLM reviewed each candidate in context and judged whether the equipment/frustration relationship was genuine. A separate LLM discovery pass then

read equipment-mentioning text without requiring a frustration keyword, because humans have been known to express irritation without help from our carefully curated vocabulary. The retained passages were additionally assigned a 1–5 comedic-value score.

2 Data and Search Corpus

The supplied project summary reports that 729 documents were scanned. All 72 initial keyword-screen hits were classified as coming from NASA oral-history material. The corpus therefore represents an oral-history search rather than a comprehensive corpus of all NASA mission records.

2.1 Equipment-side vocabulary

The equipment-side vocabulary deliberately includes both mechanical states (“stuck”, “came loose”) and retrieval/mobility states (“lost”, “floating away”, “drifting”). It also includes “stowage”, which is not itself a failure word but is relevant to the organizational problem of controlling objects.

2.2 Frustration-side vocabulary

The corresponding frustration/profanity screen included “hell”, “damn”, “shoot”, “frustrat”, “crap”, “aggravate”, and “annoy”.

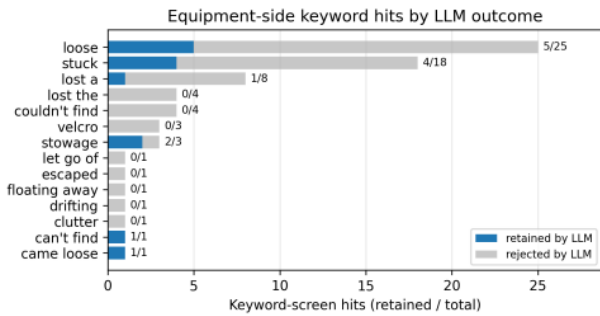


Fig. 1: Equipment-side keyword hits (all 72 screened candidates), split by LLM outcome. Labels give retained/total. These are screening hits, not unique incidents.

Keyword counts (Figures 1 and 2) are tallied *before* LLM review and are split by review outcome. They may overlap within documents or passages and should not be interpreted as independent observations or as frequencies of profanity in the corpus.

3 Methods

The implemented analysis is a multi-stage retrieval and adjudication pipeline rather than a single keyword search. The code separates

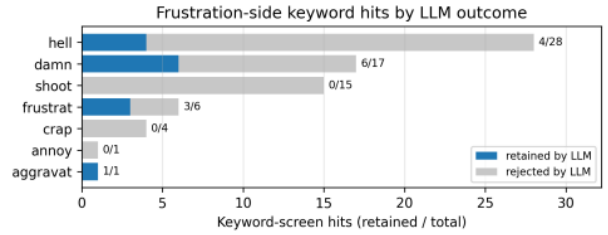


Fig. 2: Frustration/profanity keyword hits (all 72 screened candidates), split by LLM outcome. Labels give retained/total; “shoot” (15 hits) was never retained.

corpus construction, deterministic lexical retrieval, contextual LLM adjudication, and a second LLM-based discovery pass. The reported 72 candidates and 14 retained cases come from the deterministic-search/LLM-review branch, whereas the discovery branch is designed to find cases that the fixed frustration vocabulary could miss.

3.1 Stage 0: corpus construction and preservation

The first component crawls the NASA Johnson Space Center Oral History Project. The crawler follows only URLs within the configured NASA JSC oral-history scope and accepts both HTML transcript pages and PDF documents linked from those pages. HTML is converted to plain text with BeautifulSoup, while PDFs are converted with a PDF text extractor when available. Each saved document retains its source URL and basic document-type metadata. The crawler maintains a checkpoint containing visited URLs, the pending queue, and the number of saved pages, allowing an interrupted run to resume rather than restart the collection.

In parallel, the NTRS client queries NASA’s public Technical Reports Server API using configured searches for topics including ISS stowage, spacecraft habitability and clutter, crew-equipment restraint, on-orbit tool loss, and human-factors stowage. For each returned record, the pipeline saves citation metadata and, when available, full text or extracted PDF text. NTRS retrieval is therefore a cross-reference corpus rather than a replacement for the oral-history collection.

For the reported run, the supplied project summary records 729 documents scanned and 72 lexical-screen hits; all 72 hits were classified as oral-history hits. Thus, although the software supports a two-source corpus, the reported candidate set did not contain an NTRS hit.

3.2 Stage 1: deterministic lexical candidate generation

The candidate generator uses explicitly defined equipment-side and frustration-side vocabularies. The patterns are compiled as case-insensitive regular expressions. For every occurrence of an equipment-side expression, the analyzer tokenizes the document using non-whitespace spans and defines a symmetric window extending 40 words on either side of the equipment match. A candidate is emitted when at least one frustration-side expression occurs within that window. The stored record contains the source type, filename, matched equipment expression, matched frustration expression, and normalized context snippet. This procedure is deliberately a high-recall screening rule rather than a semantic classifier. It does not determine whether the two terms refer to the same event, whether the frustration was caused by the equipment problem, or whether either term is being used literally. Those questions are deferred to contextual review.

The resulting candidate file, `hits.csv`, is therefore an event-level retrieval table rather than a list of independent incidents. A single document can produce multiple hits, and multiple hits can describe the same operational episode.

3.3 Stage 2: contextual LLM adjudication

Every lexical candidate is passed to an LLM review function. The review prompt asks whether the snippet describes frustration caused by an equipment or stowage problem, rather than merely containing both vocabulary classes. For each candidate, the review records four annotations: a binary genuine-link decision, a 0–1 confidence value, a 1–5 comedic-value score, and a short textual rationale. The curated table is constructed by retaining only records for which `genuine_link=True`. The 14 retained cases therefore represent model-adjudicated positives within the 72 deterministic candidates, not manually verified ground truth.

The supplied implementation is provider-agnostic at the calling layer. The configuration defines an eight-slot endpoint pool spanning Gemini, Groq, OpenRouter, and DeepSeek, with two configured slots per provider. Requests rotate through the pool and use per-slot cooldowns and failover for transient failures. HTTP 429, 5xx, timeout, and malformed-response conditions trigger retry/cooldown behavior, whereas authentication

or model-availability errors can mark a slot unusable for the run. JSON responses are parsed strictly first and then with limited tolerance for code fences or surrounding prose.

The review results are cached using a hash of the snippet and matched keywords. Consequently, rerunning the analysis after downstream code changes need not spend additional API calls on candidates that have already been adjudicated. These engineering details improve operational robustness but do not make the LLM judgments independent ground truth.

3.4 Stage 3: independent LLM discovery scan

A separate discovery stage addresses a weakness of fixed-vocabulary retrieval. Rather than looking only at passages already containing a frustration keyword, it asks an LLM to read chunks of documents containing at least one equipment-side keyword and to identify equipment-related frustration or cursing directly from the text. Documents are divided into overlapping 600-word chunks with a 60-word overlap, with a maximum of 40 chunks per document. The discovery prompt targets explicit profanity, redacted profanity, descriptions of swearing, frustration lacking an obvious keyword, sarcasm, and paraphrased expressions.

Each discovery finding contains its source type, file, chunk index, a short quotation, a finding type, a 1–5 comedic-value annotation, and a rationale. Because adjacent chunks overlap, identical normalized quotations from the same file are deduplicated. Chunk-level results are cached using a SHA-256 hash of the filename, chunk index, and chunk text, allowing failed or interrupted runs to resume without repeating successful calls. This discovery branch is methodologically distinct from the 72-hit lexical screen. Its findings should not be added to the numerator or denominator of the 14/72 calculation unless a separate adjudication protocol is applied. In particular, the existence of this second branch means that the study has both a keyword-dependent retrieval result and an LLM-assisted discovery result; they answer related but different retrieval questions.

3.5 Outcome definitions and quantitative measures

Three quantities describe the deterministic retrieval branch. The candidate rate is

$$R_{\text{candidate}} = \frac{72}{729} = 9.88\%. \quad (1)$$

The contextual positive yield, or screening precision, is

$$P_{\text{screen}} = \frac{14}{72} = 19.4\%. \quad (2)$$

The retained-positive rate relative to the scanned document count is

$$R_{\text{retained}} = \frac{14}{729} = 1.92\%. \quad (3)$$

These quantities describe the behavior of this particular retrieval and adjudication procedure. They are not estimates of the prevalence of equipment-related frustration, or profanity in NASA operations. A prevalence estimate would require a defined sampling frame, a validated incident-level coding scheme, and an estimate of false negatives in addition to the observed false positives.

3.6 Candidate-to-Signal Ratio

To connect corpus scope, candidate burden, and contextual yield, we define a descriptive Candidate-to-Signal Ratio (CSR). Let N be the number of scanned documents, C the number of candidate hits, and G the number of LLM-adjudicated genuine links. Define

$$D = \frac{G}{N}, \quad B = 1 + \frac{C - G}{C}, \quad (4)$$

and

$$\text{CSR} = \frac{D}{B} = \frac{G/N}{1 + (C - G)/C}. \quad (5)$$

For the supplied run, $\text{CSR} = 1.064\%$. This is a constructed pipeline statistic, not a standard information-retrieval metric.

3.7 Confidence-weighted extension and calibration constraint

The raw CSV contains a continuous LLM confidence annotation $c_i \in [0, 1]$ for every candidate. It is tempting to reinterpret c_i as a posterior probability $p_i = P(\text{genuine}_i = 1 \mid \text{context})$ and replace G by $\sum_i p_i$. Such a substitution is mathematically coherent only when the confidence scores have been calibrated against an independent reference standard.

This distinction is important in the present dataset. The supplied manuscript explicitly states that the confidence values are outputs of the same model-based adjudication process and have not been calibrated against independent human labels. Thus, a value such as $c_i = 0.95$ cannot be reported as a 95% probability of correctness.

For a future calibrated version of the study, let p_i denote calibrated posterior probabilities. The expected signal mass is

$$\mu_G = \sum_{i=1}^C p_i, \quad (6)$$

and the corresponding expected CSR is

$$\text{CSR}_\mu = \frac{\mu_G/N}{1 + (C - \mu_G)/C}. \quad (7)$$

This formalizes the proposed continuous-probability metric without conflating model confidence with calibrated probability.

A second extension penalizes uncertainty using binary Shannon entropy,

$$H(p_i) = -p_i \log_2 p_i - (1 - p_i) \log_2 (1 - p_i), \quad (8)$$

with the convention $0 \log 0 = 0$. The entropy-adjusted quality is

$$\mathcal{Q}_\sigma = \frac{1}{N} \sum_{i=1}^C p_i [1 - H(p_i)]. \quad (9)$$

Under this definition, a high-confidence positive receives high weight, while a maximally uncertain $p_i = 0.5$ receives zero weight.

Observed-data diagnostic. The supplied `hits.csv` contains 72 confidence values whose sum is 66.19. The mean confidence among the 58 rejected candidates is 0.924 (median 0.95, maximum 1.00), while the mean among the 14 retained candidates is 0.900 (maximum 0.95). The rejected candidates are thus, on average, *more* confidently classified than the retained ones. This is what one would expect if c_i expresses the model's confidence in whichever decision it made, so that a confident "no" scores as high as a confident "yes." Under that reading, c_i is not $P(\text{genuine})$ even in principle, and summing it over all 72 candidates measures decisiveness rather than signal. For transparency, if the raw c_i values are inserted into the formulas above merely as an uncalibrated confidence mass, they give a raw-confidence CSR of 8.402% and an entropy-weighted confidence diagnostic of 6.208%. These are diagnostics, not estimates of expected genuine-link counts. In particular, the supplied data do not reproduce an expected-positive count of 17.01; the raw confidence sum is 66.19. A claim of $\mu_G = 17.01$, $\text{CSR}_\mu = 1.323\%$, or $\mathcal{Q}_\sigma = 1.269\%$ therefore requires an additional calibration or transformation step that is not documented in the supplied files.

3.8 Comedic-value annotation

The LLM assigns each retained candidate a 1–5 comedic-value score, with higher values intended to indicate a stronger improbable-research anecdote. This score is an interpretive annotation rather than a validated psychological or linguistic measurement. The project summary reports an average of 2.1/5 among “genuine hits.” Re-computation from `hits.csv` shows that 2.07 is the mean over all 72 candidates; the mean over the 14 retained candidates is 3.36 (range 3–4), and the mean over the 58 rejected candidates is 1.76. The summary figure therefore appears to have been averaged over the wrong set, and we report the retained-set value throughout.

3.9 Reproducibility and failure handling

The pipeline is designed for long-running public-data retrieval and LLM processing. HTTP requests use a shared session with retry logic for transient server errors and a configured politeness delay. Oral-history and NTRS acquisition are checkpointed. LLM review results and discovery chunks are cached. The LLM discovery stage can abort after repeated pool-wide failures while leaving successful chunks cached for later continuation. Command-line options allow the operator to skip scraping, rerun analysis only, or disable either LLM stage. These engineering features do not themselves establish scientific validity, but they reduce accidental differences caused by interrupted computation and make the computational procedure more reproducible.

Quantity	Value
Documents scanned	729
Keyword-screen candidates	72
LLM-confirmed genuine links	14
Distinct source documents among retained	12
Candidate rate	9.88%
Screening precision	19.4%
Retained-positive rate	1.92%
CSR	1.064%
Mean comedic value, retained ($n = 14$)	3.36
Mean comedic value, rejected ($n = 58$)	1.76

Table 1: Principal pipeline quantities.

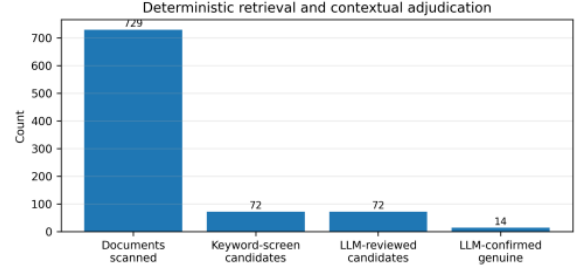


Fig. 3: Deterministic retrieval and contextual adjudication: from 729 documents to 72 keyword candidates to 14 LLM-retained links.

4 Results

4.1 Deterministic retrieval and contextual yield

The deterministic branch began with 729 scanned documents and generated 72 keyword-screen candidates. All 72 were sent to contextual LLM review, and 14 were retained as genuine equipment/frustration links (Table 1). The corresponding rates are $72/729 = 9.88\%$, $14/72 = 19.4\%$, and $14/729 = 1.92\%$. The 72 count is a count of keyword co-occurrence hits, not 72 independent incidents: the 72 hits come from 56 documents, and the 14 retained hits come from 12. Two pairs of retained hits (Hutchinson, 2000; Kranz, 1998) describe the same passage, so the retained set contains at most 12 distinct episodes. Accordingly, $72/729$ is a hit-to-document ratio and $14/729$ a retained-hit-to-document ratio, not incident prevalence measures.

4.2 Lexical profile and what the LLM review adds

The equipment-side vocabulary was dominated by “loose” (25 hits) and “stuck” (18), followed by “lost a” (8), “lost the” (4), and “couldn’t find” (4). On the frustration side, “hell” was the most frequent matched expression (28), followed by “damn” (17), “shoot” (15), “frustrat” (6), and “crap” (4).

The screen was designed to over-retrieve, so retention rate is a property of each keyword rather than a verdict on the study (Figure 5). Explicit frustration vocabulary was the most precise: “aggravat” retained 1 of 1 and “frustrat” 3 of 6 (50%), against 6 of 17 for “damn” (35%) and 4 of 28 for “hell” (14%). The low rate for “hell” is unsurprising, because “hell” is mostly idiom (“what the hell,” “all hell breaks loose” used for chaos in general) and so almost always co-occurs

with equipment words by accident. “Shoot” and “crap” retained none. On the equipment side, “stowage” retained 2 of 3, “loose” 5 of 25 (20%), and “stuck” 4 of 18 (22%). The co-occurrence matrix (Figure 4) shows where the retained links sit: they concentrate in the “loose”/“stuck” rows crossed with “damn,” “hell,” and “frustrat.” In sum, 58 of the 72 candidates (80.6%) were not retained; this is the contextual stage doing the job it was added for, and the 14 positives remain LLM-adjudicated cases, not validated ground truth.

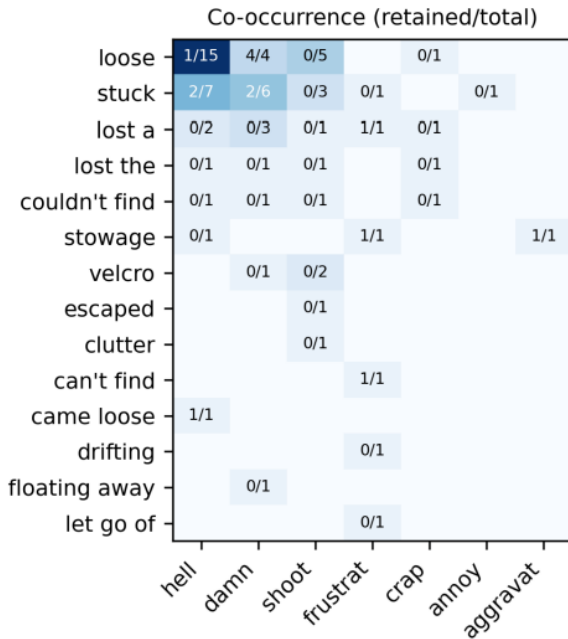


Fig. 4: Equipment-side (rows) by frustration-side (columns) co-occurrence among the 72 candidates, annotated retained/total.

4.3 Retained passages

Table 2 lists eight of the retained passages. The strongest anecdotes are those in which the language mirrors the physical problem. In Kranz’s account of an Agena thruster that was “stuck on,” the crew undocks and “all hell really broke loose.” In Hutchinson’s account of an unmanned launch, “all hell broke loose, literally,” and the sentence continues to a micrometeorite shield that “came loose” and jammed a solar panel. (The figurative phrase precedes the literal event in the transcript; the two occur within a few dozen words, and the screen cannot say more than that.) Hughes recalls tape that kept coming loose over the driver’s seat at highway speed, prompting “God damn, Deke.”

Inspection also shows that the retained set is not all gold. Several hits are idiomatic or weakly con-

nected to a frustrating equipment event: “poor old Gus damn near was killed by his first hatch, when it blew loose”; “it’s a damn good thing” that an EVA suit was durable; and “pull the damn gate loose,” where the gate belongs to a simile about a horse. Kranz’s “all hell” describes a spacecraft roll, not an emotional state. These cases are consistent with the paper’s central caution that LLM adjudication is triage rather than ground truth, and they strengthen the case for human coding.

The recurring joke is supplied by physics itself: when nothing is attached to anything, even a very small object can acquire the dramatic importance of a missing sock in a washing machine, except that the washing machine is orbiting Earth.

4.4 Score distributions

Figure 6 shows that the model’s confidence values pile up near 0.9–1.0 for both outcomes, and that the only values above 0.95 belong to rejected candidates. Comedic-value scores (Figure 8) separate the groups cleanly: retained candidates score 3 or 4, whereas rejected candidates are concentrated at 1–2, which is expected if the comedic score partly tracks how well a passage ties frustration to a physical event.

4.5 Confidence cutoffs

Figure 7 compares the hard CSR after imposing successive confidence cutoffs with a raw-confidence-mass diagnostic ($\sum c_i$ among surviving candidates). The hard CSR falls to zero above a cutoff of 0.95 because no retained candidate exceeds that value, which is a further sign that c_i describes decisiveness rather than the probability of a genuine link. The curve is a diagnostic of the score distribution, not a posterior-probability estimate.

4.6 Independent discovery branch

The independent discovery stage (Section 3.4) found 25 findings, with a mean comedic value of 3.12/5: 17 frustration-without-keyword cases, 7 explicit-profanity cases, and 1 described-swearing case. These are not merged with the principal 14/72 statistic because they arise from a different retrieval branch, and their mean score is comparable to, not higher than, that of the retained keyword set (3.36). The highest-scoring findings (5/5) are both from one Melroy interview about a microgravity toilet: the crew “tore the place apart looking for the opera gloves,” and a stowage bag was pulled “and pulling and pulling” but would not come back out. Other findings

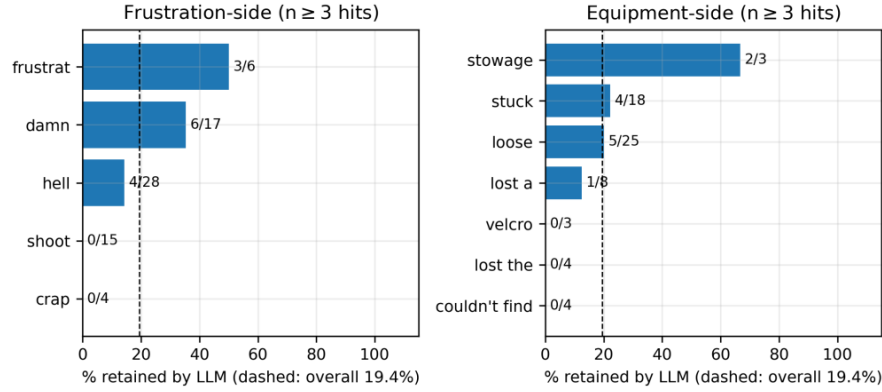


Fig. 5: Share of keyword-screen hits retained by the LLM, by keyword (keywords with at least three hits). Labels give retained/total; the dashed line is the overall 19.4%.

Source (JSC OHP)	Keywords	Score	Excerpt
Hutchinson, 5 Jun 2000	loose / hell	4	“You know, all hell broke loose, literally. . . . [the shield] came loose and then it jammed one of the solar panels”
Kranz, 19 Mar 1998	stuck / hell	4	“The thruster’s stuck on somewhere . . . when they undocked . . . all hell really broke loose”
Jeffs, 21 May 1998	loose / damn	4	Loose solder balls in zero g “essentially close the damn circuit for you”
Cernan, 11 Dec 2007	can’t find / frustrat	4	“It’s so easy to get frustrated in space. In zero gravity you can’t find anything, it floats away.”
Hughes, 10 Sep 2013	loose / damn	3	Gray tape “would come loose” over the driver’s seat at 90 mph: “God damn, Deke.”
Thompson, 29 Aug 2000	stuck / damn	3	“this damned antenna was sticking up there” (recovery craft, low electrical wires)
Hartsfield, 12 Jun 2001	stowage / aggravat	3	“it was just aggravating to have to undo all that, because that locker . . . was a really stowage issue”
McArthur, 26 Mar 2018	stowage / frustrat	3	“something I found very frustrating on Space Station, the inability to find [items]”

Table 2: Selected retained passages with the LLM’s 1–5 comedic-value score. Excerpts are abridged. Dates are interview dates inferred from file names.

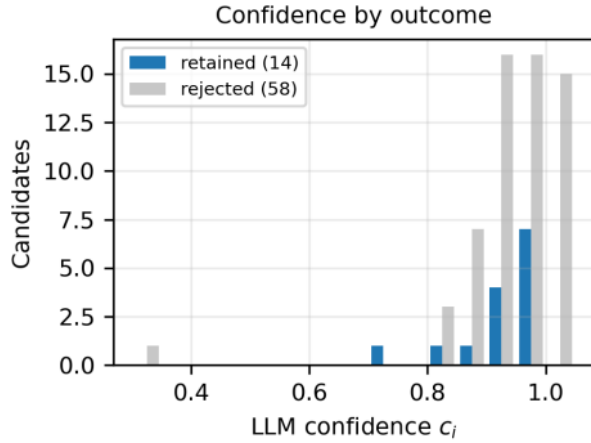


Fig. 6: Distribution of LLM confidence for retained and rejected candidates. Rejected candidates are, on average, the more confidently classified.

include “All hell breaks loose when you fire an SSME” (Kraft; 4/5), a lunar science package that falls off the carry bar and “bounces across the Moon” (Duke; 4/5), and “What the hell is going on?” about a robotic-arm master alarm and a stowage problem (Homan; 4/5). Explicit-profanity findings score higher on average (3.4) than frustration-without-keyword findings (2.9) (Figure 9). Most discovery findings come from a handful of documents (Melroy and Pohl contribute four each), so they are subject to the same unit-of-analysis caveat as the keyword hits.

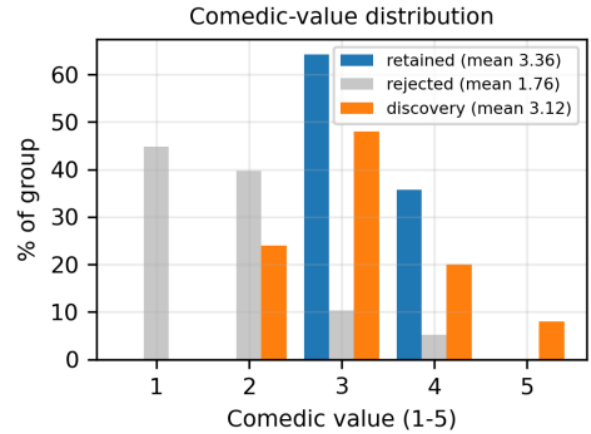


Fig. 8: Distribution of comedic-value scores (percent within each group) for retained keyword hits, rejected keyword hits, and discovery findings.

4.7 Source composition

Although the software can analyze both oral histories and NTRS technical reports, all 72 hits are oral-history material. The NTRS component demonstrates cross-reference capability but does not establish that the phenomenon occurred in the NTRS subset.

5 Discussion

The central result is modest but entertaining: a broad lexical search can recover a small collection of passages in which loss of control over equipment is linguistically adjacent to loss of composure. The computer did not discover that astronauts become angry when objects float away. It discovered something more useful for an improbable-research paper: that the archives contain unusually good sentences in which the hardware and the language fail in complementary directions.

The most conspicuous pattern is the collision between literal and figurative “loose.” The phrase works because the figurative expression mirrors the mechanical failure. This is precisely the kind

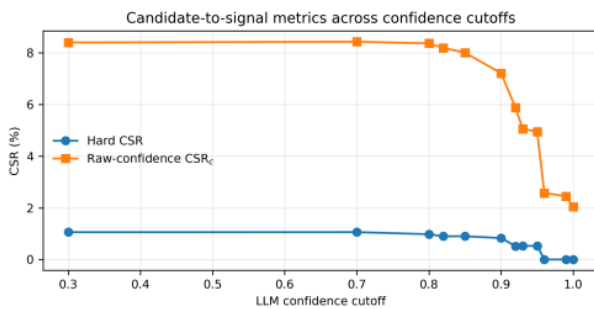


Fig. 7: Hard CSR versus a raw-confidence-mass CSR diagnostic across the distinct confidence cutoffs observed in `hits.csv`.

of low-stakes linguistic coincidence that makes the dataset suitable for an improbable-research treatment.

The results also suggest that “lost equipment” is too narrow a label for the phenomenon. Several retained cases concern equipment that is stuck, badly stowed, difficult to locate, or physically mobile. In microgravity, the operational boundary between “lost,” “floating,” and “poorly stowed” can become blurred.

Importantly, the data do not establish that equipment problems cause profanity. The LLM’s judgment is contextual classification, not experimental causal inference. The most defensible statement is that the retained passages contain contextual evidence that the frustration expression concerns the equipment problem.

5.1 The role of the LLM

The LLM is best understood as a triage and annotation instrument. It did not generate the corpus or the initial candidate set. It interpreted passages already retrieved by the lexical procedure. Broad lexical retrieval favors recall but admits many false positives, while contextual review improves specificity at the price of model-dependent judgments. A useful follow-up would be double annotation by human coders of all 72 candidates for: (1) equipment problem present; (2) frustration expression present; (3) frustration attributable to equipment; (4) profanity versus non-profanity; (5) literal/figurative lexical overlap; (6) whether multiple hits describe the same incident; and (7) comedic value. Agreement statistics would then quantify how much of the 14/72 result is reproducible independently of the LLM.

5.2 Implications for confidence-weighted retrieval

The confidence extension is potentially useful because binary decisions discard information. However, a confidence score produced by an adjudicating model is not automatically a posterior probability, and here it appears to measure decisiveness rather than $P(\text{genuine})$ (Section 3.7). The appropriate next experiment is therefore not to relabel c_i as p_i , but to elicit an explicit probability of the positive class and to collect an independently coded reference set against which to fit or evaluate a calibration mapping. Once calibrated, CSR_μ and $\mathcal{Q}_\sigma$ can be evaluated alongside thresholded retrieval measures.

6 Limitations

Corpus and sampling. The corpus is a convenience corpus built from the NASA JSC Oral History Project and researcher-configured NTRS queries; all 72 hits are from oral histories. Results cannot be generalized to all NASA records, missions, or equipment incidents. Oral histories are also retrospective narratives, and speakers may have tidied their language for the transcript.

Retrieval vocabulary and window. The deterministic branch depends on a fixed list of expressions and a 40-word whitespace-delimited window. It can miss synonyms, paraphrases, redacted expressions, and relationships spanning more than 40 words, and it can join unrelated statements. The discovery scan mitigates this but is itself an LLM search, not a validated estimate of the false-negative rate.

Unit of analysis. The principal counts are keyword hits, not unique incidents, speakers, missions, or documents. The 14 positives come from 12 documents and should not be interpreted as 14 independent operational events.

LLM adjudication and discovery. The positive set depends on model judgments that are not independent ground truth, and inspection of the retained passages (Section 4.3) shows that some are idiomatic. The implementation rotates among several providers and models, so exact model composition is part of the configuration, and caching can preserve earlier annotations after prompt changes; a controlled replication should version the prompt, model identifier, and configuration with every judgment. The discovery branch reflects prompt design, model behavior, and chunk boundaries, and its results should remain separate from the 14/72 statistic unless independently adjudicated.

Confidence and comedic value. Confidence values are model annotations and are not calibrated probabilities. Comedic value is a subjective, model-generated annotation with no validated scale properties.

Source version and extraction. NASA pages, PDFs, and NTRS records can change, and PDF text extraction can lose layout, speaker structure, or punctuation. A controlled replication should preserve a frozen archival snapshot with cryptographic hashes.

Causality. The analysis cannot establish that an equipment problem caused a speaker to swear or express frustration.

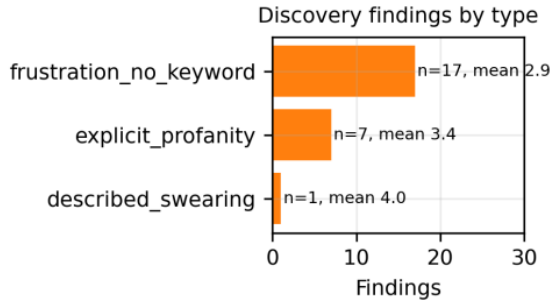


Fig. 9: Discovery-branch findings by type, with mean comedic value.

7 Recommended Reproducible Pipeline

A reproducible version of the analysis can be expressed as: (1) define the corpus; (2) define equipment and frustration vocabularies; (3) retrieve lexical co-occurrences; (4) preserve source document, local context, and keyword identity; (5) submit each candidate to a contextual classifier; (6) retain positive cases and record confidence; (7) independently inspect retained cases; and (8) report both screening yield and contextual yield. The distinction between retrieval precision and phenomenon prevalence should be maintained throughout. In particular, the 19.4% value is a screening precision, not “19.4% of NASA equipment problems involve profanity.”

8 Conclusion

This study searched 729 NASA oral-history documents for linguistic evidence connecting loose, stuck, lost, floating, or poorly stowed equipment with frustration. Seventy-two lexical candidates were reviewed, and 14 (from 12 documents) were classified as genuine equipment/frustration links, giving a 19.4% candidate-to-positive yield.

The most common screening terms were “loose,” “stuck,” and “lost,” while “hell,” “damn,” and “shoot” dominated the frustration vocabulary. The strongest anecdotes are those in which the language itself mirrors the physical problem: a thruster is “stuck on,” and all hell really breaks loose; a shield comes loose, and all hell has already broken loose, literally.

The scientific conclusion is deliberately restrained. The analysis demonstrates a searchable and reproducible association in a small oral-history corpus; it does not establish prevalence, causation, or a universal relationship between

loose hardware and profanity. In particular, 19.4% should not be read as “19.4% of space equipment causes swearing.” The confidence-weighted metrics CSR_μ and Q_σ are defined here as future calibrated metrics; the present data support only uncalibrated diagnostics.

The improbable-research conclusion is simpler: in space, apparently, even the vocabulary can lose its grip. The good news is that the language usually comes back. The bad news is that the screw-driver may not.

Data Availability

The analysis is based on the supplied project files: a screening summary, the raw keyword-hit CSV (72 candidate records with model annotations), and the LLM discovery CSV (25 findings).

Acknowledgements

The author acknowledges the NASA oral-history material underlying the search corpus and the computational tools used for candidate retrieval and contextual screening.