

# Data-Faithful Natural Language Generation for Explaining Fitness-Market Anomalies

Kauan Divino Pouso Mariano<sup>1</sup>, Fabrycio Leite Nakano Almada<sup>1</sup>  
Victor Emanuel da Silva Monteiro<sup>1</sup>, Maykon Adriell Dutra<sup>1</sup>  
Jefferson Felex de Faria<sup>2</sup>, Jennifer Vitória da Silva Peixoto<sup>2</sup>  
Kamylla Sejane Pouso Freitas<sup>3</sup>

<sup>1</sup>Institute of Informatics – Federal University of Goiás (UFG)  
Goiânia – GO – Brazil

<sup>2</sup>Polytechnic School – Pontifical Catholic University of Goiás (PUC)  
Goiânia – GO – Brazil

<sup>3</sup>School of Medicine – Alfredo Nasser University Center  
Aparecida de Goiânia – GO – Brazil

{kauan, fabrycio, victor.emanuel, maykonadriell}@discente.ufg.br

felexmc22@gmail.com, jenniferpeixoto04@gmail.com

kamyllasejane@unifan.edu.br

**Abstract.** *This study presents an integrated pipeline that transforms anomaly-detection outputs from global fitness-market data into data-faithful narratives. Using a 132-country panel and a 2019 baseline, we analyze observed post-COVID-19 recovery through 2025, while treating 2026 only as a modelled extension. Four complementary detectors—regional-median deviations, linear-regression residuals, Random Forest residuals, and Isolation Forest—are combined through method agreement and mapped to a rule-based typology. A deterministic, template-based Natural Language Generation (NLG) component then produces short, medium, and complete explanations from tabular evidence. In 2025, 14 countries were consensus anomalies; Guyana exemplified market growth without participatory expansion, with revenue recovery of +401.43%, participation recovery of -2.06%, and a 403.50-point gap. All 132 narratives passed automatic factual-consistency checks. These checks validate field preservation rather than fluency, usefulness, or explanatory quality; human evaluation remains future work.*

## 1. Introduction

COVID-19 disrupted physical activity through restrictions on mobility, gym access, and daily routines [Hasson et al. 2022]. Recovery remained uneven across populations and contexts [Eime et al. 2024, Strain et al. 2024]; although the fitness sector adapted and resumed growth, its expansion did not necessarily yield proportional gains in population participation [Cecere and Bernardi 2023, Liu et al. 2022].

This heterogeneity creates an analytical challenge for global fitness-market data, which combine market, participation, socioeconomic, health, and recovery indicators.

Rankings or descriptive statistics may hide relevant deviations, especially when economic recovery diverges from participation trends. This work investigates how anomaly-detection outputs from global fitness-market indicators can be transformed into scientifically grounded explanatory narratives. The proposed pipeline integrates tabular anomaly detection with NLG: anomaly detection identifies countries whose recovery profiles deviate from regional, contextual, or multivariate expectations, and controlled templates convert these signals into data-faithful narratives.

The approach is grounded in data-to-text generation, explainable analysis, and anomaly explanation. Data-to-text research emphasizes that NLG from structured data requires content selection, ordering, interpretation, and factual consistency evaluation, not only fluent verbalization [Lin et al. 2024]. As related work, LIDA provides a conceptual reference for multi-stage data-communication pipelines based on semantic summarization, goal identification, and faithful generation, but it was not used in our implementation [Dibia 2023]; data-storytelling research further highlights the narrative organization of data communication [Ren et al. 2023]. In parallel, Explainable Artificial Intelligence and outlier-explanation studies argue that anomalous observations need interpretable descriptions of the variables responsible for their deviation [Mersha et al. 2024, Tritscher et al. 2023, Panjei et al. 2022]. Rather than proposing a new anomaly-detection algorithm, the methodological contribution integrates four detectors, method-agreement consensus, a rule-based typology, controlled NLG, and factual-consistency checks to communicate post-COVID-19 divergences among market recovery, population participation, and health-related indicators.

## 2. Methodology

The analysis uses `data/clean_gym_data.csv`, a cleaned country-year panel derived from the Kaggle World Gym & Fitness Trends (2000–2026) dataset and covering 132 countries.<sup>1</sup> Its documentation lists the Health & Fitness Association, World Health Organization Global Health Observatory, World Bank Open Data, EuropeActive/Deloitte, and Our World in Data as institutional sources; however, this study did not reconstruct the primary collection or harmonization process. The panel combines market and socioeconomic or health indicators, later aggregated into country-level recovery features; global participation is summarized as a simple, unweighted mean across countries. The main analysis compares 2022–2025 with the 2019 pre-COVID-19 baseline, whereas 2026 is treated as a modelled extension rather than a consolidated observation. Cross-country comparisons may therefore reflect unresolved differences in provenance, harmonization, inflation adjustment, and currency conversion.

The pipeline follows five compact stages. First, revenue recovery, participation recovery, and `market_health_gap` are computed relative to 2019; the gap is defined as revenue recovery minus participation recovery, with positive values indicating stronger economic recovery than participation recovery and negative values indicating the opposite. This indicator is not treated as a direct public-health measure, but as an interpretable contrast linking anomaly detection to narrative explanation. Second, anomaly detection combines four complementary methods: regional-median deviations, linear-regression residuals, Random Forest residuals [Li and van Leeuwen 2023], and Isolation

---

<sup>1</sup>Implementation scripts are available at <https://github.com/kauandivino/globalfit-anomalies>.

Forest [Agyemang 2024]. Third, each method flags the top 10% of countries, with the number of cases rounded up. Method scores are z-standardized and averaged with equal weights into a combined score, while agreement is counted separately. A consensus anomaly is flagged by at least two methods; this operational rule reduces dependence on a single detector but is not a universal statistical threshold.

Fourth, a rule-based typology maps each deviation to an interpretable anomaly category and guides content selection. Finally, a deterministic, template-based NLG component produces short, medium, and complete narratives. Content realization is cumulative, as summarized in Table 1: each length preserves the preceding evidence and adds contextual slots populated directly from the anomaly and feature tables. The generator uses neither a Large Language Model nor free-form generation; predefined clauses and tabular slots constrain its output to explicit evidence. Automatic factual-consistency checks verify country and anomaly-type mentions, rounded numerical values, gap values, and absence of predefined causal terms, following data-to-text principles of content planning and factual consistency [Thomson et al. 2023, Alonso and Agirre 2024].

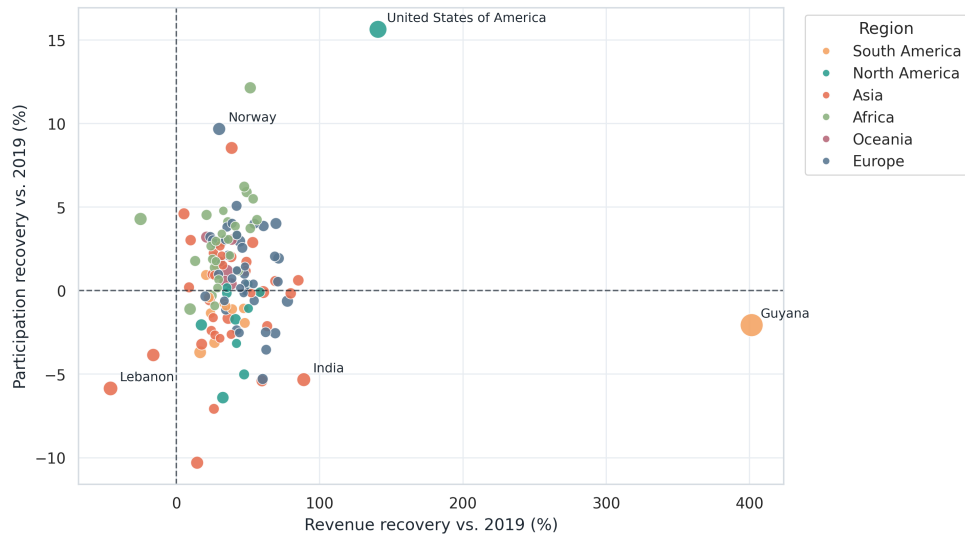
**Table 1. Incremental content plan for the three deterministic NLG variants.**

| Variant         | Required evidence slots                                                                       | Interpretive and contextual content                                                     |
|-----------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <b>Short</b>    | Country, anomaly type, revenue recovery, participation recovery, and market-participation gap | Core interpretation selected by the rule-based typology                                 |
| <b>Medium</b>   | All short-version slots                                                                       | Region and deviations from the regional median and model-based expectation              |
| <b>Complete</b> | All medium-version slots                                                                      | Selected 2019 socioeconomic and health indicators, plus the model-based reference value |

### 3. Results

Between 2019 and 2025, global health-club revenue increased by 69.97%, the number of gyms by 19.79%, and memberships by 26.21%, whereas the simple country mean of fitness participation rose only 0.38 percentage points, from 40.24% to 40.62%. This asymmetry shows that market recovery and population participation are related but not equivalent dimensions.

The anomaly-detection stage identified 14 consensus anomalies in 2025, four of which were flagged by all four methods. Residual-model validation was weak: participation obtained  $R^2$  values of approximately 0.13–0.18, while revenue and `market_health_gap` had negative  $R^2$  values; residuals are therefore treated as exploratory expected-behavior references rather than strong predictions. Sensitivity tests yielded 6, 14, and 19 consensus anomalies at 5%, 10%, and 15% thresholds, respectively. Guyana, the United States of America, India, and Lebanon remained consensus anomalies across all three settings, indicating a stable core but a threshold-dependent complete set. Figure 1 shows the 2025 relation between revenue and participation recovery, while Table 2 summarizes selected profiles.



**Figure 1. Revenue versus participation recovery by country in 2025. Colors denote regions, bubble size reflects the combined anomaly score, and labels identify selected consensus anomalies.**

**Table 2. Selected 2025 anomaly profiles; recovery is relative to 2019 and gap is in percentage points.**

| Country | Type                                                      | Agreement | Rev. (%) | Part. (%) | Gap (p.p.) |
|---------|-----------------------------------------------------------|-----------|----------|-----------|------------|
| Guyana  | Market growth without participatory expansion             | 4/4       | +401.43  | -2.06     | 403.50     |
| Lebanon | Market recovery below regional pattern                    | 4/4       | -45.95   | -5.85     | -40.10     |
| Norway  | Participation above expectation with low revenue recovery | 2/4       | +29.85   | +9.68     | 20.17      |

The agreement column reports how many of the four detectors flagged each country: Guyana and Lebanon represent unanimous 4/4 agreement, whereas Norway illustrates the minimum 2/4 operational consensus. The table also shows the sequence from evidence to explanation. For Guyana, revenue recovery of +401.43%, participation recovery of -2.06%, and a 403.50-point gap triggered the type *market growth without participatory expansion*. The corresponding corrected English template rendering is: “*Guyana showed market growth without participatory expansion in 2025: fitness-sector revenue increased by 401.43% relative to 2019, participation decreased by 2.06%, and the market-participation gap reached 403.50 percentage points. This profile represents strong sectoral expansion without proportional population engagement.*” Thus, the rule-based label selects the interpretation, while tabular values provide its evidence. Because relative changes may be amplified by small initial values, the extreme percentage is treated as an anomaly signal rather than externally validated national evidence.

In a 2025 pipeline run, all 132 narratives passed the automatic factual-consistency checks, with a mean score of 1.0 for country and anomaly-type mentions, rounded recovery and gap values, and absence of predefined causal terms. These checks validate template filling and factual consistency, but do not assess fluency, clarity, usefulness, or

explanatory quality, and therefore do not replace human evaluation.

#### 4. Discussion and Final Remarks

The applied contribution integrates complementary anomaly detectors, consensus, rule-based typology, deterministic NLG, and factual-consistency checks, making explanations traceable to explicit indicators. Fitness-market recovery should not be equated with public-health improvement [Duijvestijn et al. 2023]. However, the analysis identifies associations rather than causal effects; residual models are exploratory, the complete anomaly set is threshold-sensitive, and cross-country interpretation remains limited by unresolved provenance and harmonization issues.

Template constraints do not demonstrate lower hallucination rates than alternative generators, and a perfect factual-consistency score verifies selected inputs rather than overall textual quality. Future work should add human evaluation [Thomson et al. 2024], compare templates with Large Language Models and neural or free-form generation, assess unsupported content, externally validate national indicators, improve expected-value models, and test other domains.

#### Artificial Intelligence Usage

Generative artificial intelligence (AI) tools were used during manuscript preparation to support language revision, technical documentation, LaTeX editing, and presentation. Specifically, Prism, ChatGPT, and Claude, including reasoning-enabled modes, were used to paraphrase excerpts, improve English clarity, draft and revise LaTeX table structures, review textual consistency with reported outputs, and refine figure captions. These tools were editorial and technical aids only; they were not part of the implemented NLG pipeline and did not generate the narratives evaluated in this study. They did not determine the research questions, analytical methods, empirical results, or scientific interpretations. All AI-assisted content was reviewed and validated by the authors, who assume full responsibility for the manuscript.

#### References

- Agyemang, E. F. (2024). Anomaly detection using unsupervised machine learning algorithms: A simulation study. *Scientific African*, 26:e02386.
- Alonso, I. and Agirre, E. (2024). Automatic logical forms improve fidelity in table-to-text generation. *Expert Systems with Applications*, 238:121869.
- Cecere, R. and Bernardi, P. (2023). From crisis to opportunity: exploring the interplay between entrepreneurial processes and technological changes in italian fitness industry. *International Journal of Entrepreneurship and Innovation Management*, 27(3/4):237–268.
- Dibia, V. (2023). LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In Bollegala, D., Huang, R., and Ritter, A., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, Toronto, Canada. Association for Computational Linguistics.

- Duijvestijn, M., de Wit, G. A., van Gils, P. F., and Wendel-Vos, G. C. W. (2023). Impact of physical activity on healthcare costs: a systematic review. *BMC Health Services Research*, 23(1):572.
- Eime, R., Harvey, J., and Charity, M. (2024). Australian sport and physical activity behaviours pre, during and post-covid-19. *BMC Public Health*, 24(1):834.
- Hasson, R., Sallis, J. F., Coleman, N., Kaushal, N., Nocera, V. G., and Keith, N. (2022). Covid-19: Implications for physical activity, health disparities, and health equity. *American Journal of Lifestyle Medicine*, 16(4):420–433.
- Li, Z. and van Leeuwen, M. (2023). Explainable contextual anomaly detection using quantile regression forests. *Data Mining and Knowledge Discovery*, 37(6):2517–2563.
- Lin, Y., Ruan, T., Liu, J., and Wang, H. (2024). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.
- Liu, R., Menhas, R., Dai, J., Saqib, Z. A., and Peng, X. (2022). Fitness apps, live streaming workout classes, and virtual reality fitness for physical activity during the covid-19 lockdown: An empirical study. *Frontiers in Public Health*, Volume 10 - 2022.
- Mersha, M., Lam, K., Wood, J., AlShami, A. K., and Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111.
- Panji, E., Gruenwald, L., Leal, E., Nguyen, C., and Silvia, S. (2022). A survey on outlier explanations. *The VLDB Journal*, 31(5):977–1008.
- Ren, P., Wang, Y., and Zhao, F. (2023). Re-understanding of data storytelling tools from a narrative perspective. *Visual Intelligence*, 1(1):11.
- Strain, T., Flaxman, S., Guthold, R., Semenova, E., Cowan, M., Riley, L. M., et al. (2024). National, regional, and global trends in insufficient physical activity among adults from 2000 to 2022: a pooled analysis of 507 population-based surveys with 5.7 million participants. *The Lancet Global Health*, 12(8):e1232–e1243.
- Thomson, C., Reiter, E., and Belz, A. (2024). Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*, 50(2):795–805.
- Thomson, C., Reiter, E., and Sundararajan, B. (2023). Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.
- Tritscher, J., Krause, A., and Hotho, A. (2023). Feature relevance xai in anomaly detection: Reviewing approaches and challenges. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.