LEXARA-RF: Reference-Free Metrics for Evaluating Conversational Visual Analytics Agents

Srishti Palani* Vidya Setlur*

Tableau Research, Palo Alto, CA, USA

ABSTRACT

Conversational visual analytics (CVA) agents powered by large language models generate visualizations and natural-language explanations from open-ended queries. Evaluating these multimodal outputs is challenging: curated reference benchmarks are costly to author, cannot comprehensively capture the space of valid responses, and are unavailable in production. Building on the Lexara evaluation framework, we introduce LEXARA-RF, a reference-free set of metrics that scores CVA outputs using only the prompt, data, and model response. We reformulate evaluation as verification: 13 metrics operationalize visualization design theory and Gricean cooperative principles as computable consistency, intent-alignment, and design validity checks. On a human-rated corpus of CVA test-cases, LEXARA-RF achieves alignment comparable to reference-based formulations, outperforms surface-similarity NLG baselines, and localizes structurally grounded failures with high accuracy.

Index Terms: Conversational Visual Analytics, Evaluation Metrics, Visualization Design, Cooperative Communication Principles.

1 INTRODUCTION

Advances in large language models (LLMs) have enabled conversational visual analytics (CVA) systems that allow users to generate and refine visualizations through natural language [29, 30]. While these systems lower barriers to data exploration, evaluating their outputs remains a challenge. CVA agents produce multimodal outputs, i.e., data-grounded visualizations paired with natural language explanations that must be assessed for data fidelity, encoding effectiveness, and alignment with user intent across a conversation.

Existing evaluation approaches [15, 4, 23] compare model outputs against curated references (i.e., acceptable responses for a given input). However, the space of valid CVA responses is inherently open-ended: a single query may admit multiple valid visualizations differing in chart type, encoding, or aggregation, making exhaustive coverage infeasible and introducing false negatives [7]. Moreover, constructing and maintaining references is costly, requiring expert effort to enumerate acceptable outputs and update them as data and tasks evolve [15, 4]. In production settings, references are often unavailable, as user queries and data continuously shift, preventing evaluation at inference time and scalable monitoring of CVA agent quality [28, 3, 11].

To address these limitations, we propose LEXARA-RF, a reference-free extension of the Lexara evaluation framework [23] that assesses CVA outputs as intrinsic artifacts grounded in the user prompt and underlying data. We reformulate evaluation as verification: each metric tests whether a response satisfies conditions derived from visualization design theory [16, 6] and conversational principles [8, 30], rather than whether it matches a curated reference output. We contribute: (1) 13 reference-free metrics spanning *expressiveness*, *effectiveness*, and *conversational alignment*, with full pseudocode and judge prompts; (2) a validation on a human-rated CVA corpus (N = 120) showing parity with reference-based metrics at this sample size and substantially higher alignment than surface-similarity NLG baselines (BLEU, ROUGE-L, BERTScore); and (3) evidence that the metrics localize structurally grounded failure modes while surfacing semantically ambiguous ones as distributed signals, supporting hybrid evaluation. Together, these contributions establish reference-free evaluation as a practical complement to reference-based benchmarks for CVA, enabling continuous integration, production monitoring, and rapid prompt iteration where reference authoring is prohibitive.

*E-mails: {srishti.palani, vsetlur}@salesforce.com.

2 RELATED WORK

This work builds on three research themes: (1) theoretical foundations for visualization design, (2) automated visualization evaluation, and (3) conversational quality in analytical dialogue.

2.1 Visualization Design Theory

A rich body of work distinguishes between visualization expressiveness (encoding all and only the intended facts) and effectiveness (leveraging perceptual channels for accurate interpretation) [16, 6]. Prior research further formalizes visual encoding channels and their perceptual properties, alongside principles of graphical integrity that emphasize truthful and efficient data representation [2, 36]. More recent frameworks organize visualization design across levels from data abstraction to interaction, outlining how systems support analytical workflows [20, 31]. These insights have been operationalized in systems such as Show Me and Draco, which generate visualizations using heuristics and constraint-based reasoning [17, 19]. Unlike these generative systems, LEXARA-RF operationalizes the same principles for evaluation, transforming them into reference-free metrics that verify whether generated visualizations satisfy established design constraints.

2.2 Automated Visualization Evaluation

Benchmarks such as nvBench [15], VisEval [4], and NLVCorpus [33] evaluate models against curated reference outputs, while Lexara [23] extends this paradigm with graded metrics spanning visualization and natural language quality. More recent work explores richer evaluation pipelines by decomposing assessments across representation levels, training learned critics from human examples, or using latent similarity representations [26, 24, 13, 32]. However, all ultimately depend on reference visualizations, either explicitly or implicitly via training data. This dependence incurs high annotation costs, penalizes valid alternative responses, and precludes evaluation in production where references are unavailable. Parallel reference-free work in NLP [14, 18, 37, 10, 9] evaluates text without references but do not address the multimodal, design-sensitive nature of CVA. LEXARA-RF bridges this gap by grounding reference-free evaluation in data, visualization design principles, and conversational alignment.

2.3 Conversational Quality for Analytical Dialogue

CVA systems also produce natural language explanations that must be evaluated for communication quality. Grice's cooperative prin-

ciples (Quality, Quantity, Relation, and Manner) provide a foundation for assessing truthfulness, informativeness, relevance, and clarity [8]. Prior work adapts these principles to analytical dialogue, emphasizing factual consistency between text and visualization, context preservation across turns, coherence, and insightfulness [35, 30, 5, 21]. These frameworks are inherently reference-free, evaluating responses based on internal consistency and conversational context rather than comparison to expected outputs. However, they have not been systematically operationalized as computable metrics for CVA systems. LEXARA-RF instantiates these principles through verification by checking consistency across text, visualization, data, and conversation history, enabling end-to-end reference-free evaluation. Unlike learned critics [24, 26] and reference-free NLG evaluators [14, 18, 37, 10, 9], which rely on supervision from reference exemplars, LEXARA-RF requires no reference data at training or evaluation time and grounds its metrics directly in visualization theory and conversational principles.

3 LEXARA-RF: REFERENCE-FREE EVALUATION METRICS

We propose LEXARA-RF, a set of reference-free metrics that evaluate CVA outputs as intrinsic artifacts grounded in the user prompt and underlying data source, rather than curated references. LEXARA-RF operates on the following runtime inputs: the user utterance U, visualization specification V and natural language explanation NL, underlying data source and schema (D,S), and conversation history H. While aligned with prior CVA evaluation dimensions, our metrics derive signals directly from visualization theory and conversational principles instead of curated references. They are implemented as verification procedures that test consistency, intent alignment, and design validity, producing graded scores on a common 0–100 scale. Full pseudocode for each metric and additional worked examples spanning aggregation, filter, sort, encoding, and multi-turn failure modes are provided in the paper's Supplementary Materials. For reproducibility, all implementation choices were fixed a priori: semantic similarity uses `sentence-transformers/all-MiniLM-L6-v2` embeddings with cosine similarity [27, 38]; fuzzy string matches use `RapidFuzz` token-set ratios normalized to $[0,1]$ [1]; metric outputs are normalized to the shared $[0,100]$ scale; thresholds include 0.3 for Data Fidelity plausibility, ≥ 0.5 for filter field/value matching, and $1.5 \times$ IQR for outlier detection; and full 13-metric evaluation requires approximately 18 seconds per response, or about USD 5 in API charges for the 120-response corpus based on the OpenAI API pricing schedule [22], with details in the Supplementary Materials.

We employ fixed weights and penalties while scoring the metrics to reflect the relative importance and reliability of different signals, grounded in visualization theory and verification strength rather than learned from data. The weights, 13 metric dimensions, and 0–100 scoring scale are inherited from Lexara [23], preserving interpretability and enabling direct comparison between reference-based and reference-free formulations (§4.2). Signals verified directly against the data receive the highest weight as objective evidence; structural and semantic checks receive moderate weight; multiplicative penalties capture errors with compounding perceptual impact (e.g., axis misuse).

We organize our reference-free metrics into three themes: *expressiveness*, *effectiveness*, and *conversational alignment*. Rather than comparing responses to references, the metrics verify consistency with visualization theory and conversational principles.

3.1 Expressiveness of the Visualization

Expressiveness captures whether a visualization encodes the intended data and analytical intent. We verify this by checking that the specification V is consistent with the data (D,S) and aligned with the user query U.

Data Fidelity verifies that the chart's values are computationally correct. We compute three signals: (1) s_1 (structural validity) checks that fields exist in the schema, operations match data types, and fields plausibly correspond to concepts extracted from U; (2) s_2 (pipeline re-execution) independently executes V's transformation pipeline on D and compares the result to the chart output, treating the specification as the claim and the data as the evidence. Alternative aggregations (e.g., `SUM` vs. `COUNT`) are evaluated to detect mislabeling; (3) s_3 (intent alignment) verifies that the fields and aggregations in V match the analytical intent inferred from U. If $s_1 < 50$, the specification is structurally invalid and the score is s_1; otherwise, the score is $0.2\,s_1 + 0.5\,s_2 + 0.3\,s_3$, weighting re-execution highest as the strongest correctness signal.

Field Similarity verifies that the visualization binds the fields implied by the user's query. We extract analytical concepts from U (e.g., "revenue," "category," "last quarter") and match each concept to schema columns in S. For each concept, we score candidate fields using a weighted combination of semantic similarity, fuzzy string matching, and type compatibility, selecting the highest-scoring column as the best match. The final score measures how many of these best-matched fields are actually used in V. This approach captures semantic alignment rather than exact matching: a query for "*revenue by category*" should map to fields like `Sales` and `Category`, while a chart using unrelated fields such as `Discount` or `Region` scores low despite being syntactically valid.

Filter Accuracy verifies that the visualization applies the constraints implied by the user's query. We extract temporal, categorical, and numeric filter intents from U, augmenting relevant filters from prior turns using conversation history H. We normalize V's filters and match them to these intents using semantic similarity over fields and values. Scoring uses an asymmetric penalty that prioritizes missing constraints: $100 \cdot m/|F_U| - 25 \cdot \text{miss} - 15 \cdot \text{extra}$, where missing filters are penalized more than unrequested additions. This reflects that omitted constraints violate user intent, while extra filters only narrow the result. Only explicit and persistent filters are evaluated; implicit data source defaults are ignored.

Sort Accuracy verifies that the visualization respects ordering implied by the user's query. We detect sort signals in U, including superlatives (e.g., "top," "highest") and explicit directives (e.g., "sort by X"), and infer the sort field and direction. If no sort is requested, the metric returns 100 when V is unsorted. Otherwise, we compare the inferred sort to V's specification. Scoring combines field and direction alignment: $100 \cdot s_f \cdot d$, where s_f measures similarity between expected and actual sort fields, and $d \in 1.0, 0.5$ rewards correct or reversed direction. Missing required sorting scores 0.

3.2 Effectiveness of the Visualization

Effectiveness captures whether a visualization uses perceptual channels appropriately for accurate interpretation, matching encodings to the data and analytical task [16, 6, 2].

Chart Type verifies that the chosen mark type aligns with the user's intent and data characteristics. We combine two signals: (1) intent inferred from U (e.g., trend, comparison, distribution, correlation) to identify expected chart types, and (2) Show Me [17], which produces a ranked set of appropriate marks from the inferred fields. The score is 100 if V matches the top recommendation, 50 if it appears in the candidate set, and 0 otherwise, with a small bonus for matching the inferred intent.

Axis Accuracy verifies that axis assignments support accurate interpretation. Since position is the most precise perceptual channel [6], incorrect assignments can distort meaning. We audit key conventions with multiplicative penalties: (a) temporal intent requires encoding time on a positional channel; (b) time should appear on the x-axis when both axes are present; (c) bar charts should place measures on the y-axis and categories on x; (d) bar and area charts should use a zero baseline; and (e) explicit axis directives in

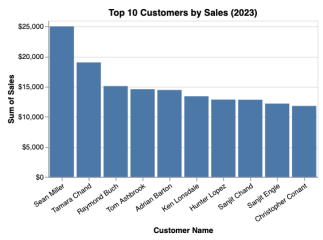
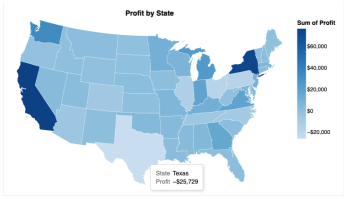
Test	User Utterance	Model's CVA Response (V, NL)	Computed LEXARA-RF Metrics	Interpretability of Metrics
1	`show top 10 customers by profit in 2023`	 *"These are the most profitable customers in 2023."*	Data Fidelity = 92% ; Field Similarity = 50% ; Sort Accuracy = 70% ; Filter Accuracy = 100% ; Chart Type = 100% ; Axis Accuracy = 100% ; Visual Encoding = 100% ; Interactivity = 100% . **Overall Viz** 89% . Factual Grounding = 30% ; Assumptions Disclosure = 25% ; Insightfulness = 40% ; Coherence = 90% . **Overall NL** 46% .	**Expressiveness and Conversational Alignment failures: field swap.** V picks `Sales` where U asked for `Profit` affecting *Field Similarity*, *Sort Accuracy*, and *Data Fidelity* scores. *Conversational Alignment* surfaces the downstream consequence: *Factual Grounding* and *Insightfulness* drop because NL's "most profitable" claim is incorrect; *Assumptions Disclosure* drops because the silent `Profit` to `Sales` substitution is never flagged.
2	`show profit by state`	 *"Profit varies across the United States, with California leading the way."*	Data Fidelity = 100% ; Field Similarity = 100% ; Filter Accuracy = 100% ; Sort Accuracy = 100% ; Chart Type = 100% ; Axis Accuracy = 100% ; Visual Encoding = 40% ; Interactivity = 100% . **Overall Viz** 93% . Factual Grounding = 70% ; Assumptions Disclosure = 50% ; Insightfulness = 30% ; Coherence = 90% . **Overall NL** 60% .	**Effectiveness and Conversation Alignment failures: encoding mismatch.** Sequential single-hue palette on signed `Profit` (range −$25K to $76K) renders profit losses visually indistinguishable from low-profit states, affecting *Visual Encoding*. *Insightfulness*, *Assumptions Disclosure* drop because this obscuring palette choice is not articulated. *Expressiveness* encodes correct fields and marks.

Table 1: Worked examples showing how LEXARA-RF reference-free metrics evaluate CVA responses and help diagnose failures.

U must be respected. Violations incur multiplicative penalties to reflect their compounding effect on interpretability.

Visual Encoding verifies that non-positional channels (color, shape, size, opacity, text) are used appropriately. Perceptual principles recommend matching channels to data types (e.g., hue for categories, size for magnitude) [2, 19]. We score each channel using *practice* (alignment between channel and data type), and *relevance* (whether the encoded field is implied by U). The final score is the mean of 0.6· practice and 0.4· relevance across used channels. Mismatched encodings reduce interpretability, while irrelevant fields introduce visual clutter.

Interactivity verifies that tooltips provide relevant details on demand [31]. Required fields include those referenced in U and those defining the visualization (e.g., axes, color). We compute two signals: *coverage*, the fraction of required fields present, and *usability*, which penalizes duplicate fields and missing units. The final score is 0.7· coverage +0.3· usability, capturing both completeness and clarity.

3.3 Conversational Alignment

CVA responses include natural language explanations that must support analytical reasoning and interpretation. We ground evaluation in Grice's cooperative principles, adapted for analytical conversation, which assess utterances in context rather than against references [8, 30].

Factual Grounding (Quality) verifies claims in NL against V and recomputed values from D, penalizing contradictions more heavily than omissions.

Assumptions Disclosure (Quantity) checks whether implicit choices in V (e.g., mappings, filters, aggregations) are surfaced in NL. Programmatically extracted assumptions are scored by an LLM judge, since undisclosed assumptions cannot enter shared context [5].

Insightfulness extends Quantity by assessing whether NL surfaces meaningful patterns (e.g., trends, outliers), grounded in summary statistics from D and V [21].

Coherence (Manner) evaluates clarity and logical structure of NL.

Follow-up Relevance (Relation) assesses multi-turn alignment using conversation history H, verifying persistence of filters, analytical scope, and entity references [5].

Overall Viz and **Overall NL** are the mean of their constituent metrics, with per-metric scores reported for interpretability.

4 Validation of Reference-Free CVA Metrics

We evaluate LEXARA-RF along three dimensions: *Concurrent Validity*: alignment with expert human judgments; *Convergent Validity*: agreement with reference-based evaluation despite not requiring curated expected outputs, and discriminant separation from surface-similarity NLG baselines; and *Diagnostic Utility*: localization of interpretable failure modes via per-metric score decomposition. All evaluations reuse the human-rated CVA response corpus from Palani & Setlur [23]: $N = 120$ responses stratified by metric score ranges, ambiguity types (syntactic, semantic, pragmatic), and analytical task types (descriptive, comparative, trend), drawn from a generative corpus of evaluation experiments across eight LLMs (GPT-5 family, o3, o4-mini, Claude Opus 4, Claude 3.7 Sonnet, DeepSeek R1), six system prompts, and analyst-supplied data sources spanning finance, education and healthcare domains. Each response carries expert human ratings on a 0–100 scale per metric from two Lexara-trained CVA practitioners (median linear-weighted Cohen's $\kappa = 0.65$ for viz metrics, 0.63 for NL metrics, per [23]). Because of the sample size, we present this as a descriptive validation rather than an equivalence test; per-metric results are provided in the Supplementary Materials.

4.1 Concurrent Validity: Human Alignment

All metrics are computed solely from (U, V, NL, D, S, H) without reference outputs. Human scores are the mean of two expert ratings. Inter-rater agreement was substantial, with a median linear-weighted Cohen's $\kappa = 0.78$ (IQR 0.71–0.83) across metrics and corpus-wide $\kappa = 0.76$, comparable to prior CVA evaluation on the same corpus [23].

Per-metric correlations are reported in the Supplementary Materials, with aggregate values of Spearman's $\rho_{\text{med}} = 0.60$ (range 0.42–0.74 across metrics) and Kendall's $\tau_{b\text{med}} = 0.47$ (across metrics). Deterministic, execution-grounded metrics (Data Fidelity, Filter Accuracy, Sort Accuracy, Factual Grounding) show the strongest alignment with human judgment ($\rho = 0.68$–0.74),

whereas rubric-based LLM-as-Judge metrics (Insightfulness, Coherence) exhibit lower but consistent agreement ($\rho = 0.42$–0.52), reflecting the greater subjectivity of interpretive evaluation [21, 23]. This distinction also provides practical guidance on which metrics are suitable for automated gating versus human review.

4.2 Convergent Validity: Reference-Based Alignment

We compare each reference-free metric against (i) its reference-based counterpart from Palani & Setlur [23] and (ii) standard NLG baselines (BLEU [25], ROUGE-L [12], BERTScore [39]), all evaluated on the same response set. Agreement with human judgment is assessed using Steiger's Z-test [34] for dependent correlations, with Holm–Bonferroni correction across the 13 comparisons.

Reference-free metrics fail to reject parity with reference-based formulations on 11 of 13 dimensions ($p > 0.05$ after correction), with mean $|\Delta\rho| = 0.04 \pm 0.01$ on the remaining two dimensions. Because $N = 120$ provides limited power to detect small effects, this result should be interpreted as indistinguishable performance at the current sample size, not strict measurement equivalence; detecting $\Delta\rho = 0.05$ with 80% power would require approximately $N \approx 400$. These findings support reference-free metrics as a practical complement to reference-based evaluation when reference authoring is prohibitively expensive.

NLG baselines. BLEU, ROUGE-L, and BERTScore correlate with human judgment at $\rho = 0.18 \pm 0.02$, 0.24 ± 0.03, and 0.46 ± 0.03, respectively; substantially below reference-free and reference-based formulations (see Supplementary Materials). This indicates that surface similarity to a reference *NL* string does not capture the structurally grounded quality of CVA responses, and that the parity result above is non-trivial: any metric correlating at $\rho \approx 0.60$ with humans would already be doing more than a strong text-similarity baseline.

Divergent case analysis. We analyzed cases where reference-free and reference-based scores differed by more than 0.15 (90th percentile of pilot-phase score variation), yielding 36 responses. Two annotators independently coded each case (Cohen's $\kappa = 0.74$) as valid alternatives, phrasing-invariant disclosures, or true errors, resolving disagreements through discussion [23, 30]. The distribution was 44%, 32%, and 24%, respectively. Although descriptive of this subset only, the low proportion of true errors suggests that most divergences reflect legitimate under-specification in CVA tasks rather than failures of the reference-free metrics.

4.3 Diagnostic Utility: Failure Localization

To examine whether individual metrics surface their target failure modes, we analyzed a stratified subset of $N = 90$ responses from the Lexara corpus [23] across performance terciles (high, medium, low) and the three metric pillars. Per-mode counts range from 5 to 11 cases (Supp C Table 2) and reflect the natural distribution of failure types in the source corpus rather than enforced balance; we treat the structurally grounded subset (54 responses across six modes) as our primary basis for localization claims. Two Lexara-trained annotators independently labeled each response using an 11-category failure taxonomy adapted from Lexara (Cohen's $\kappa = 0.79$), with disagreements resolved through discussion (Supplementary Materials). The taxonomy was derived independently of the metric implementations in Section 3 and comprised:

- **Expressiveness:** incorrect data values, field mapping, aggregation, filters, and sorting.
- **Effectiveness:** inappropriate chart type, axis errors, and visual encoding mismatches.
- **Conversational Alignment**: factual grounding failures, undisclosed implicit assumptions, and multi-turn context loss.

For each labeled response, we identify the metric with the largest negative standardized deviation from its corpus-level distribution and treat it as the primary diagnostic signal, while assessing whether deviations are localized or distributed. Two examples illustrate this behavior. T47 omitted a year filter from the preceding turn produces a sharp drop in Filter Accuracy while other metrics remain near baseline, cleanly isolating the failure. The failure to disclose an implicit population-normalization assumption in T63 produces coordinated declines in Assumptions Disclosure, Insightfulness, and Coherence, reflecting a conversational failure.

Across the corpus, structurally grounded failures (e.g., missing filters, aggregation errors, baseline violations) produce localized deviations in their corresponding metrics, whereas semantically ambiguous failures (e.g., undisclosed assumptions, pragmatic interpretation) produce correlated declines across conversational alignment metrics. Thus, the framework localizes well-defined structural errors while surfacing ambiguity as distributed signals. Per-failure-mode breakdown in the Supplementary Materials.

5 Discussion: Limitations & Future Work

These results support a measured claim: LEXARA-RF aligns with expert judgment at correlations comparable to reference-based methods (median $\rho = 0.60$), exceeds surface-similarity NLG baselines, and localizes structurally grounded failure modes at $\geq 90\%$ accuracy without curated references. At $N = 120$, observed parity reflects limited statistical power rather than strict equivalence; accordingly, we position LEXARA-RF as a complement to reference-based evaluation that eliminates per-test-case ground-truth authoring. Rubric-based metrics show lower agreement and weaker localization, warranting different gating thresholds and continued human review. Three factors plausibly explain why rubric-based metrics underperform: (i) inherent subjectivity in interpretive tasks like Insightfulness and Coherence [21]; (ii) LLM-as-Judge variance even at temperature 0 (details in supplementary materials); and (iii) scope overlap among Insightfulness, Coherence, and Assumptions Disclosure, which manifests as correlated rather than localized signals when failures are semantically ambiguous (§4.3). Promising improvement paths include ensemble judging across model families, calibration of rubric anchors against held-out human ratings, and decomposing broader rubrics into narrower verification primitives. Only 24% of reference-free vs. reference-based disagreements (§4.2) are true metric errors, whereas 44% are *valid alternatives*, suggesting much of what reference-based evaluation flags is legitimate under-specification rather than metric failure. Per-metric decomposition effectively localized failure dimensions in our case studies, suggesting promise for regression triage and prompt debugging in CVA pipelines.

Three limitations bound our claims: First, the framework is validated only on Vega-Lite-style specifications; extending metrics that inspect encodings or interaction requires adaptation to other grammars (e.g., ggplot, D3). Second, evaluation is limited to English utterances and judge prompts. Third, validation uses the Lexara corpus with inherited metric weights; broader datasets spanning grammars, languages, and analytical tasks are needed for data-driven calibration. Future work should target a larger validation corpus spanning multiple grammars and languages, data-driven weight calibration, and extension to richer CVA tasks such as multi-table joins and calculated fields. Our results highlight a hybrid evaluation paradigm: automated, reference-free metrics like LEXARA-RF can reliably handle structurally grounded correctness checks, while human judgment remains essential for interpretive and subjective aspects of analytical communication.

Data Availability. The 120-response validation corpus is available at https://github.com/SrishtiPalani/Lexara-CVA-Eval/tree/main/lexara_rf_corpus. Supplementary materials have all details for reproducibility.

REFERENCES

[1] M. Bachmann. RapidFuzz: Rapid Fuzzy String Matching in Python and C++. https://github.com/rapidfuzz/RapidFuzz, 2025. Accessed 2026-06-24. 2

[2] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, WI, 1983. ISBN: 978-0299090609. 1, 2, 3

[3] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley. The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. In *IEEE International Conference on Machine Learning and Systems*, 2017. doi: 10.1109/BigData.2017.8258038 1

[4] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311, 2025. doi: 10.1109/TVCG.2024.3456320 1

[5] H. H. Clark and S. E. Brennan. Grounding in Communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, eds., *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006 2, 3

[6] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080 1, 2

[7] S. Fu, K. Xiong, X. Ge, S. Tang, W. Chen, and Y. Wu. Quda: Natural Language Queries for Visual Data Analytics. 2020. doi: 10.48550/arXiv.2005.03257 1

[8] H. P. Grice. Logic and Conversation. In P. Cole and J. L. Morgan, eds., *Syntax and Semantics, Vol. 3: Speech Acts*, pp. 41–58. Academic Press, New York, 1975. ISBN: 978-0126235500. 1, 2, 3

[9] T. S. Kim, H. Lee, Y. Lee, and J. Kim. Evalet: Evaluating Large Language Models through Functional Fragmentation. In *Conference on Human Factors in Computing Systems*. ACM, 2026. doi: 10.1145/3772318.3790285 1, 2

[10] W. Kryscinski, B. McCann, C. Xiong, and R. Socher. Evaluating the Factual Consistency of Abstractive Text Summarization. In B. Webber, T. Cohn, Y. He, and Y. Liu, eds., *ENLP*, pp. 9332–9346, 2020. doi: 10.18653/v1/2020.emnlp-main.750 1, 2

[11] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2011. doi: 10.1109/TVCG.2011.279 1

[12] C.-Y. Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. 4

[13] C. Liu, C. Da, X. Long, Y. Yang, Y. Zhang, and Y. Wang. SimVecVis: A Dataset for Enhancing MLLMs in Visualization Understanding. *arXiv preprint arXiv:2506.21319*, 2025. 1

[14] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In H. Bouamor, J. Pino, and K. Bali, eds., *ENLP*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153 1, 2

[15] T. Luo, C. Huang, L. Shen, B. Li, S. Shen, W. Zeng, N. Tang, and Y. Luo. nvBench 2.0: A Benchmark for Natural Language to Visualization under Ambiguity. 2025. doi: 10.48550/arXiv.2503.12880 1

[16] J. Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions on Graphics*, 5(2):110–141, 1986. doi: 10.1145/22949.22950 1, 2

[17] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594 1, 2

[18] P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *ENLP*, 2023. doi: 10.18653/v1/2023.emnlp-main.557 1, 2

[19] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 2019. doi: 10.1109/TVCG.2018.2865240 1, 3

[20] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111 1

[21] C. North. Toward Measuring Visualization Insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. doi: 10.1109/MCG.2006.70 2, 3, 4

[22] OpenAI. Openai api pricing. https://openai.com/api/pricing/, 2026. Accessed 2026-06-24. 2

[23] S. Palani and V. Setlur. Lexara: A User-Centered Toolkit for Evaluating Large Language Models for Conversational Visual Analytics. In *Conference on Human Factors in Computing Systems*. ACM, 2026. doi: 10.1145/3772318.3790735 1, 2, 3, 4

[24] B. Pan, Y. Fu, K. Wang, J. Lu, L. Pan, Z. Qian, Y. Chen, G. Wang, Y. Zhou, L. Zheng, Y. Tang, Z. Wen, Y. Wu, J. Lu, B. Zhu, M. Zhu, B. Zhang, and W. Chen. VIS-Shepherd: Constructing Critic for LLM-based Data Visualization Generation. 2025. doi: 10.48550/arXiv.2506.13326 1, 2

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. p. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135 4

[26] L. Podo, M. Ishmal, and M. Angelini. Vi(E)va LLM! A Conceptual Stack for Evaluating and Interpreting Generative AI-based Visualizations. 2024. doi: 10.48550/arXiv.2402.02167 1, 2

[27] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *ENLP*, 2019. 2

[28] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28, 2015. 1

[29] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A Natural Language Interface for Visual Analysis. In *ACM Symposium on User Interface Software and Technology*, pp. 365–377, 2016. doi: 10.1145/2984511.2984588 1

[30] V. Setlur and M. Tory. How Do You Converse with an Analytical Chatbot? Revisiting Gricean Maxims for Designing Analytical Conversational Behavior. In *Conference on Human Factors in Computing Systems*, pp. 1–17, 2022. doi: 10.1145/3491102.3501972 1, 2, 3, 4

[31] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, pp. 336–343, 1996. doi: 10.1109/VL.1996.545307 1, 3

[32] Y. Song, X. Zhao, and R. C.-W. Wong. Marrying Dialogue Systems with Data Visualization: Interactive Data Visualization Generation from Natural Language Conversations. In *Conference on Knowledge Discovery and Data Mining*, pp. 2733–2744, 2024. doi: 10.1145/3637528.3671935 1

[33] A. Srinivasan, N. Nyapathy, B. Lee, S. M. Drucker, and J. Stasko. Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations. In *Conference on Human Factors in Computing Systems*, pp. 1–10, 2021. doi: 10.1145/3411764.3445400 1

[34] J. H. Steiger. Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin*, 87(2):245, 1980. doi: 10.1037/0033-2909.87.2.245 4

[35] M. Tory and V. Setlur. Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation. In *Visual Analytics Science and Technology*, pp. 93–103. IEEE, 2019. doi: 10.1109/VAST47406.2019.8986918 2

[36] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983. ISBN: 978-0961392147. 1

[37] A. Wang, K. Cho, and M. Lewis. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. pp. 5008–5020. ACL, 2020. doi: 10.18653/v1/2020.acl-main.450 1, 2

[38] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, vol. 33, 2020. doi: 10.5555/3495724.3496209 2

[39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTscore: Evaluating Text Generation with BERT. 2019. doi: 10.48550/arXiv.1904.09675 4