Annals of Science and Inquiry
Vol. 2 No. 6 (2026)

Which Screenshot Goes with the Dialogue Turn? Reference Association in Human Browser Replays

Blake King[1] · Noah Knight[2] · Benjamin Lewis[3]

[1] Slovenia · [2] Slovakia · [3] Malta

Published: June 17, 2026 | Article 850

Abstract

Dialogue and browser events do not always carry the same state fields. Using all 100 validation demonstrations in a fixed WebLINX release, we compare three ways to attach existing screenshot references to dialogue turns. The 1,670 dialogue records lack direct screenshot references; 1,130 support all three rules. Within that common set, adding a preceding-time restriction to replay carry-forward changes the referenced asset in 3.01% of turns. Timestamp-nearest matching changes it in 43.10% and selects later-clock browser records in 44.60%. Its mean absolute timestamp distance is 20.18 seconds smaller than the restricted rule's, an improvement guaranteed in direction by its optimization objective. Unmatched prefixes and repeated references remain explicit. These findings concern associations between released records, without establishing image capture times, visual correctness or agent performance.

Which Screenshot Goes with the Dialogue Turn? Reference Association in Human Browser Replays

Abstract

Dialogue and browser events do not always carry the same state fields. Using all 100 validation demonstrations in a fixed WebLINX release, we compare three ways to attach existing screenshot references to dialogue turns. The 1,670 dialogue records lack direct screenshot references; 1,130 support all three rules. Within that common set, adding a preceding-time restriction to replay carry-forward changes the referenced asset in 3.01% of turns. Timestamp-nearest matching changes it in 43.10% and selects later-clock browser records in 44.60%. Its mean absolute timestamp distance is 20.18 seconds smaller than the restricted rule's, an improvement guaranteed in direction by its optimization objective. Unmatched prefixes and repeated references remain explicit. These findings concern associations between released records, without establishing image capture times, visual correctness or agent performance.

1. A state reference needs an association rule

A conversational browsing record can alternate between a person speaking and a person acting in a browser. A screenshot may accompany the browser action while the intervening utterance contains only dialogue information. Preparing a multimodal history then requires a decision: which existing browser-state reference should accompany that utterance? Choosing the preceding record and choosing the closest timestamp sound similar, yet they impose different restrictions. An association rule becomes part of the represented history before any model receives it.

WebLINX records conversational website navigation through human demonstrations (Lù et al., 2024). Its dialogue component makes the association problem concrete rather than hypothetical. Work on grounding in communication explains why coordinating a shared context is a substantive communication problem (Clark and Brennan, 1991). Our metadata cannot determine what either participant understood. It can, however, reveal whether two preparation rules would attach the same file reference to the same dialogue turn, which is a narrower and directly inspectable question.

Interactive-agent research gives state representation practical importance. ReAct interleaves reasoning and external actions (Yao et al., 2022); Mind2Web supplies crowdsourced action sequences for real websites (Deng et al., 2023). VisualWebArena evaluates visually grounded web tasks (Koh et al., 2024), while OSWorld provides execution-based evaluation in real computer environments (Xie et al., 2024). Those settings require their own execution and evaluation evidence. The present observations are performed human actions from one release, and the analysis introduces no new agent or success evaluation.

There is also a specific distinction between an asset reference and an acquired observation. DeepVoyager-VL's Section 3.3 separates discovered image references from explicit image loading and preserves observations across interaction turns (Zhang et al., 2026). That interface distinction motivates documenting how a reference enters a trajectory. We use it as methodological background, without implementing the framework or testing whether visual evidence drives subsequent reasoning. A file handle in a replay supplies evidence of a linkage, not evidence that a model loaded the file.

Our question is therefore about a reproducible transformation: across a fixed set of dialogue records, how much do replay carry-forward, carry-forward restricted by the recorded clock, and timestamp-nearest matching change the attached screenshot reference and its time distance? The main contrasts use identical

supported turns. A separate accounting of missing prefixes explains why the total number of matched records cannot itself establish which association is preferable.

2. Order, time and the meaning of an observation

Source array position and timestamp value describe different aspects of a recorded interaction. Lamport's analysis of event ordering distinguishes clock representations from relations among events (Lamport, 1978). Allen's temporal framework likewise makes temporal relations explicit instead of reducing them to an unspecified notion of closeness (Allen, 1983). These ideas motivate retaining both fields. They do not justify reconstructing a physical event order from this release, whose array has undergone curation.

The distinction is familiar in data systems. The Dataflow Model treats out-of-order data and event-time processing as explicit design concerns (Akidau et al., 2015). We do not identify WebLINX's two fields with the event-time and processing-time clocks of that system. Here one coordinate is curated replay index and the other is a relative turn timestamp. Separately locating a screenshot's actual capture would require another observation not supplied by the retained technical schema.

Event-log research provides a reason to audit such representations before downstream analysis. Process mining builds analyses from recorded events (van der Aalst, 2016), and event-log imperfection patterns systematize problems that can arise in those records (Suriadi et al., 2017). Our audit does not diagnose the present source as erroneous. The WebLINX paper's collection appendix describes asynchronous recording, ordering corrections and screenshot realignment. A reversal in the released turn timestamps is consequently an observed property of a curated representation, rather than proof of an incorrect human action.

Measurement interpretation constrains the outcomes we can name. Operationalization connects observable fields to a proposed construct (Jacobs and Wallach, 2021); numerical operations inherit the meaning of their measurement scale (Stevens, 1946). We calculate seconds between two recorded turn coordinates and exact equality of two reference strings. Validating an inference about useful visual context would require additional evidence, consistent with the broader view of validity developed by Messick (Messick, 1995). Neither small timestamp distance nor reference agreement is a correctness key.

Partially observable decision models distinguish observations from underlying environment states (Kaelbling et al., 1998). The same conceptual caution is useful here even though we fit no decision model. A screenshot reference may repeat across multiple action records, and different reference names need not imply different pixels. We therefore distinguish a reference instance at one browser record, a source asset identifier within a demonstration, and a physical screen state. Only the first two are represented in our computations.

3. A fixed release with a minimal retained projection

The source is the official WebLINX-full repository at revision 36ef9f79b43df50e25b7f3b68e5c9f6ccf4160e8 (Lù et al., n.d.). Its official valid split lists 100 unique demonstrations. All 100 corresponding replay objects were read completely, totaling 213,001,412 bytes. Each calculated raw SHA256 matches its Large File Storage SHA256 in the fixed-revision paths metadata. Length, URL, ETag and capture time are recorded per object. These checks support object identity; they do not independently validate demonstration content.

Before freezing the full analysis, we inspected the first three lexicographically sorted validation IDs for schema and structural support. Their preliminary association counts were visible. The subsequent protocol fixes all 100 IDs, three rules, tie treatment, denominators and deletion summaries. This is a

source-inspected secondary analysis, with all three inspected demonstrations included. It is neither preregistered nor evaluated on an untouched test set. No demonstrations were added or removed in response to the eventual disagreement results.

Each replay was parsed in memory and immediately reduced to technical fields. The retained projection contains demonstration and turn keys, original array index, relative timestamp and its type status, action intent, numeric event/tab/viewport fields, and digests of screenshot and page references. Utterances, speaker fields, URLs, element HTML, action arguments and typed or copied text were discarded. We requested no image, video, page or profile assets. A hash preserves equality of a source reference string; it does not provide an anonymization guarantee.

This preservation strategy separates provenance from semantic verification. Data-provenance research distinguishes tracing a derived result to its contributing source information (Buneman et al., 2001). Our saved protocol, input-object hashes and projection hashes support that tracing. Dataset documentation also calls for making collection and intended-use conditions visible (Gebru et al., 2021). The source data card specifies CC BY-NC-SA 4.0 and additional third-party and lawful-use conditions. The original pinned terms and author attribution accompany the local projection; the code repository's license is not substituted for the data license.

4. Three fully specified association rules

Let a query be a source record of type chat with no nonempty direct screenshot reference. Let a candidate be a browser record from the same demonstration with a nonempty screenshot reference. Pairing requires finite, real, non-Boolean timestamps for both records. Missing, nonfinite or other time values would retain their status and have no computed gap. They would not receive a timestamp of zero. An unmatched association likewise has a missing reference and missing distance, not a distance of zero. The treatment avoids confusing absence with a measured value, a central distinction in missing-data analysis (Little and Rubin, 2019).

Rule R chooses the candidate with the greatest source index strictly below the query index. This is replay-order carry-forward, corresponding to the preceding-reference search documented in the official library. Rule C applies the same index maximization after also requiring candidate time to be no later than query time. C therefore imposes both recorded precedence conditions; it does not choose the latest timestamp over the entire demonstration. Rule N minimizes absolute timestamp difference over all candidates, permitting later indices and later timestamps. An exact distance tie is broken by the smaller source index. Times are compared at their parsed precision before display rounding.

Write the signed gap as candidate timestamp minus query timestamp. Positive values mean that the selected browser record has a later clock value, negative values an earlier one. The absolute gap measures distance between those two coordinates. Asset disagreement is one when the selected screenshot-reference digests differ and zero when they agree. Repeated references remain separate candidate instances during selection; the asset comparison subsequently recognizes their equality. The output keeps one row per chat query, with each rule's selected index, reference identity and gap fields. This explicitly identifies the observational unit, following the data-organization principle discussed by Wickham (Wickham, 2014).

The main denominator contains queries where R, C and N all exist. Because every C candidate is also an R and N candidate, this intersection equals C's support. We compare rules on these same rows. A second table reports availability on all finite queries. The symmetric nearest rule's later-time selections are additionally divided by whether any earlier-clock reference exists anywhere in the demonstration and

whether any earlier-index reference exists. These conditions differ when order and time disagree; neither is silently used as a substitute for the other.

Figure 1 shows this deterministic research model. Its arrows describe projection, selection and comparison, without a causal or learned component. We report proportions, means and linear-interpolated empirical quantiles, specifying the quantile convention because software alternatives can disagree (Hyndman and Fan, 1996). For every main contrast, deleting each of the 100 demonstrations produces a deterministic range. These are sensitivity summaries for the finite release. They are not confidence intervals, independent experimental repetitions or a power calculation.

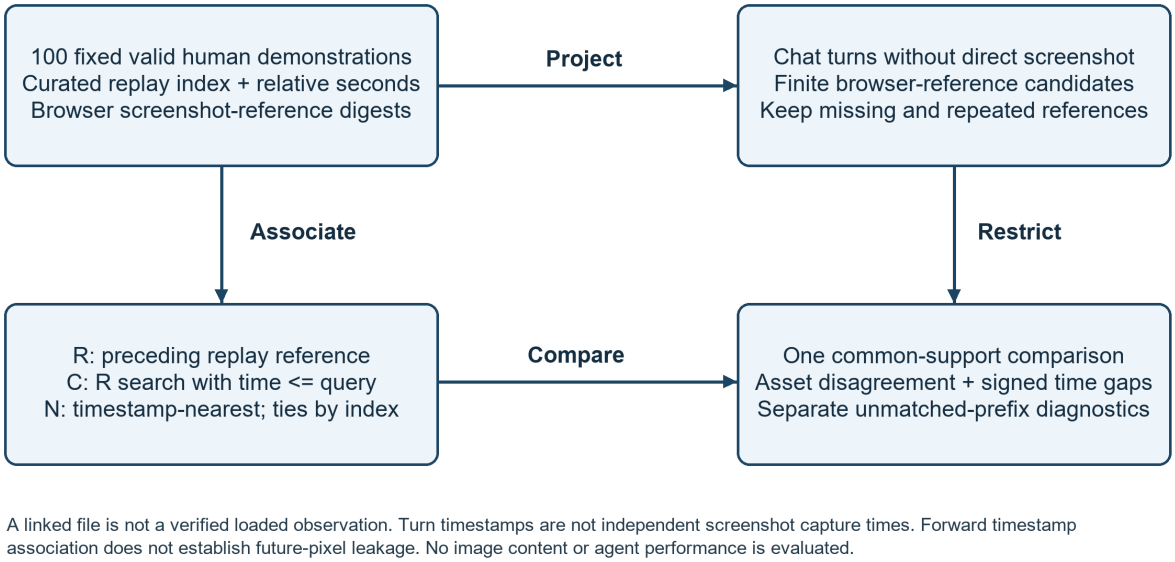

Figure 1. Reference association uses both released coordinates while retaining unmatched queries. No pixel content, loaded-observation status or successful action is inferred.

5. Support and the observed contrasts

The 100 replays contain 4,076 records: 2,406 browser turns and 1,670 chat turns. Every retained turn timestamp is finite. No chat turn directly carries a screenshot reference. Browser turns provide 2,105 screenshot-reference instances and 1,861 different reference strings counted within demonstrations. The difference records repeated links, without establishing duplicated pixels or redundant information. All sources have the published status checked; that field is not used to select turns or certify task completion. Sixty adjacent finite timestamp reversals occur across 49 demonstrations.

Table 1 separates support from reference choice. R supports 1,133 queries and C supports 1,130, whereas N supports all 1,670. Thus the added clock restriction removes three otherwise available carry-forward associations. The main set contains 1,130 queries distributed across all 100 demonstrations. The 540 queries outside that common set remain visible in the output. Assigning every one a nearest reference would answer a broader matching question, not recover missing preceding history.

Table 1. Association support among all 1,670 finite chat queries. Unmatched distances remain missing.

Rule	Available	Unmatched	Finite queries
R	1133	537	1670
C	1130	540	1670
N	1670	0	1670

Table 2 gives the paired effects on the common set. R and C select different asset references in 34 of 1,130 turns, or 3.0088%. R and N differ in 487 turns, or 43.0973%. The restricted carry-forward change is consequently much smaller in this release than the change to symmetric nearest matching. This is an empirical difference in magnitude; neither disagreement percentage measures a rate of wrong visual context. In 36 common-set turns R selects a record with a later timestamp. Two of those cases retain the same asset identity under C, illustrating why instance timing and asset disagreement are separate outcomes.

Table 2. Finite paired effects on the same 1,130 queries. Ranges delete one whole demonstration at a time; shares are on a zero-to-one scale and are not population estimates.

Contrast	Estimate	Leave-one-demo range	N
Different asset reference, R versus C	0.030088	[0.025067, 0.030825]	1130
Different asset reference, R versus N	0.430973	[0.427419, 0.435210]	1130
Later reference-record time under R	0.031858	[0.026858, 0.032638]	1130
Later reference-record time under N	0.446018	[0.441150, 0.450402]	1130
Mean absolute gap difference, N minus C (s)	-20.179210	[-20.401893, -18.151636]	1130

N selects a later-clock reference record in 504 common-set turns, or 44.6018%; C has zero such selections by construction. Mean absolute gap is 39.5563 seconds under C and 19.3771 under N. Their paired mean difference is -20.1792 seconds, with deletion values from -20.4019 to -18.1516 seconds. The nonpositive direction follows from N minimizing distance over a set that includes C's selection. Its observed magnitude adds information, but a hypothesis test of that guaranteed direction would add none. The time values describe paired records and cannot be interpreted as screenshot capture delay.

Table 3. Absolute reference-record timestamp gaps in seconds, restricted to the common set. Later-time counts use signed gaps; every row has N=1,130.

Rule	Mean absolute gap	Median absolute gap	95th percentile	Later-time count
R	38.1251	20.4580	124.4663	36
C	39.5563	21.0650	135.8180	0
N	19.3771	10.4130	72.0617	504

On all finite queries, N makes 1,041 later-time associations. Table 4 separates the support boundaries behind that larger count. In 535 cases no browser reference anywhere in the demonstration has a timestamp at or before the query. Of those, three nevertheless have an earlier-index reference, reflecting the distinction between coordinates. Another two later-time selections have an earlier-clock candidate somewhere but no earlier-index candidate. The 504 cases with both kinds of preceding support coincide

numerically with the common-set later-time count here; that equality is observed, not an assumption for other releases. No nearest-distance ties occur in this cohort, although the tie rule remains specified.

Table 4. Support categories within the 1,041 later-time N associations. These categories describe reference records, without an independent image capture clock.

Any reference at or before query time?	Any earlier-index reference?	Count
No	No	532
No	Yes	3
Yes	No	2
Yes	Yes	504

6. What the matching differences establish

The contrast demonstrates that time-nearest preparation and sequential carry-forward are materially different transformations in this source. Their common-set agreement cannot be inferred from the presence of timestamps alone. The small R-C disagreement also deserves preservation: enforcing both precedence coordinates changes only a limited portion of available asset links. Presenting only the larger all-query nearest count would obscure both that small effect and the substantial number of unmatched prefixes.

Fitness for a purpose is the relevant quality question. Wang and Strong distinguish data quality from accuracy alone (Wang and Strong, 1996); Pipino et al. develop assessment around explicit dimensions (Pipino et al., 2002); Batini and Scannapieco connect quality assessment to data and information processes (Batini and Scannapieco, 2016). Here the purpose is assembling an indexed observation history. Preserving two coordinates, missing links and reference identity makes that transformation inspectable. The analysis does not provide a quality ranking of demonstrations or a recommendation that C should replace every other association convention.

Timestamp direction also has a narrower meaning than model leakage. Leakage research concerns information paths that invalidate a claimed predictive evaluation (Kapoor and Narayanan, 2023). We train no predictor and observe no independent moment when screenshot pixels became available. A later browser-record timestamp can therefore establish only a relation between record coordinates. Confirming a prospective information violation would require capture or loading timestamps, the actual policy inputs and an evaluation design with a stated information cutoff. Those missing ingredients cannot be supplied by relabeling a positive gap.

The release bounds generalization in another way. Digital-trace analyses inherit selection and collection mechanisms (Olteanu et al., 2019), and digital social research requires aligning available traces with the question being asked (Salganik, 2017). The official validation split is a fixed benchmark partition rather than a probability sample of browsing conversations. Repeated turns from one demonstration share their recorded interaction context. Whole-demo deletion ranges describe this dependence through composition changes, while leaving claims about other websites, participants, releases and systems unresolved.

The retained files support numerical reproducibility without retaining conversational content. Inspectable computation is a distinct research objective (Peng, 2011), and saving intermediate decisions and provenance facilitates it (Sandve et al., 2013). The accompanying outputs preserve every query, rule selection and missing gap, plus per-demo counts and raw-object verification records. An independent implementation reconstructs the selections and summaries from the minimal source projections. These

checks establish computational consistency; they do not restore discarded images or replace substantive validation of visual context.

7. Conclusion

In this complete 100-demonstration validation snapshot, choosing a screenshot reference for a dialogue turn is a consequential preparation decision. Adding the recorded-time restriction to replay carry-forward changes 34 asset links on the common set; switching to nearest-time matching changes 487. The latter achieves its distance objective while often choosing later-clock browser records. Its numerical advantage therefore describes a different matching rule, rather than demonstrated improvement in perception or task execution.

A useful trajectory representation should preserve what made an association possible. In these records, source index, relative timestamp, repeated asset identity and unmatched prefix jointly determine that boundary. Reporting them together permits a reader to reproduce the transformation and choose a convention appropriate to a later study. Establishing what participants or models actually saw would require additional observation evidence beyond the reference metadata analyzed here.

References

1. Suriadi, S.; Andrews, R.; ter Hofstede, A.H.M.; Wynn, M.T.. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information Systems. https://doi.org/10.1016/j.is.2016.07.011
2. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677
3. Zhang, Huanyao; Zhou, Jiepeng; Zhao, Runhao; Shan, Yanzhe; Chen, Jiaoyang; Zhou, Bowen; Li, Bo; Wang, Fang; Wu, Jialong; Tao, Zhengwei; Mei, Lang; Yu, Xiaohan; Liu, Liyan; Chen, Chong; Zhang, Wentao. (2026). DeepVoyager-VL: Incentivizing Vision-in-the-Loop Search for Long-Horizon Multimodal Agents. arXiv:2608.01827. https://doi.org/10.48550/arXiv.2608.01827.
4. Buneman, Peter; Khanna, Sanjeev; Tan, Wang-Chiew. (2001). Why and Where: A Characterization of Data Provenance. Database Theory - ICDT 2001, 316-330. https://doi.org/10.1007/3-540-44503-X_20.
5. Deng, Xiang; Gu, Yu; Zheng, Boyuan; Chen, Shijie; Stevens, Samuel; Wang, Boshi; Sun, Huan; Su, Yu. (2023). Mind2Web: Towards a Generalist Agent for the Web. arXiv preprint. https://arxiv.org/abs/2306.06070
6. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
7. Sandve, Geir Kjetil; Nekrutenko, Anton; Taylor, James; Hovig, Eivind. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1003285
8. Akidau, Tyler; Bradshaw, Robert; Chambers, Craig; Chernyak, Slava; Fernández-Moctezuma, Rafael J.; Lax, Reuven; McVeety, Sam; Mills, Daniel; Perry, Frances; Schmidt, Eric; Whittle, Sam. (2015). The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. Proceedings of the VLDB Endowment. https://doi.org/10.14778/2824032.2824076
9. Salganik, Matthew J.. (2017). Bit by Bit: Social Research in the Digital Age. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691158648/bit-by-bit
10. Wickham, Hadley. (2014). Tidy Data. Journal of Statistical Software. https://doi.org/10.18637/jss.v059.i10
11. Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.. American Psychologist. https://doi.org/10.1037/0003-066x.50.9.741
12. Little, Roderick; Rubin, Donald. (2019). Statistical Analysis with Missing Data, Third Edition. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9781119482260
13. Batini, Carlo; Scannapieco, Monica. (2016). Data and Information Quality: Dimensions, Principles and Techniques. Data-Centric Systems and Applications. https://doi.org/10.1007/978-3-319-24106-7
14. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804

15. Olteanu, Alexandra; Castillo, Carlos; Diaz, Fernando; Kıcıman, Emre. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data. https://doi.org/10.3389/fdata.2019.00013

16. Pipino, Leo L.; Lee, Yang W.; Wang, Richard Y.. (2002). Data quality assessment. Communications of the ACM. https://doi.org/10.1145/505248.506010

17. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

18. Kaelbling, Leslie Pack; Littman, Michael L.; Cassandra, Anthony R.. (1998). Planning and acting in partially observable stochastic domains. Artificial Intelligence. https://doi.org/10.1016/s0004-3702(98)00023-x

19. Clark, Herbert H.; Brennan, Susan E.. (1991). Grounding in communication.. Perspectives on socially shared cognition.. https://doi.org/10.1037/10096-006

20. Hyndman, Rob J.; Fan, Yanan. (1996). Sample Quantiles in Statistical Packages. The American Statistician. https://doi.org/10.1080/00031305.1996.10473566

21. Lamport, Leslie. (1978). Time, clocks, and the ordering of events in a distributed system. Communications of the ACM. https://doi.org/10.1145/359545.359563

22. Lù, Xing Han; Kasner, Zdeněk; Reddy, Siva. (2024). WebLINX: Real-World Website Navigation with Multi-Turn Dialogue. arXiv preprint. https://arxiv.org/abs/2402.05930

23. Koh, Jing Yu; Lo, Robert; Jang, Lawrence; Duvvur, Vikram; Lim, Ming Chong; Huang, Po-Yu; Neubig, Graham; Zhou, Shuyan; Salakhutdinov, Ruslan; Fried, Daniel. (2024). VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. arXiv preprint. https://arxiv.org/abs/2401.13649

24. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901

25. Xie, Tianbao; Zhang, Danyang; Chen, Jixuan; Li, Xiaochuan; Zhao, Siheng; Cao, Ruisheng; Hua, Toh Jing; Cheng, Zhoujun; Shin, Dongchan; Lei, Fangyu; Liu, Yitao; Xu, Yiheng; Zhou, Shuyan; Savarese, Silvio; Xiong, Caiming; Zhong, Victor; Yu, Tao. (2024). OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. arXiv preprint. https://arxiv.org/abs/2404.07972

26. Wang, Richard Y.; Strong, Diane M.. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. https://doi.org/10.1080/07421222.1996.11518099

27. Yao, Shunyu; Zhao, Jeffrey; Yu, Dian; Du, Nan; Shafran, Izhak; Narasimhan, Karthik; Cao, Yuan. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint. https://arxiv.org/abs/2210.03629

28. Allen, James F.. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM. https://doi.org/10.1145/182.358434

29. van der Aalst, Wil. (2016). Process Mining: Data Science in Action. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

30. Lù, Xing Han; Kasner, Zdeněk; Reddy, Siva. (n.d.). WebLINX-full [Human browser demonstration dataset, revision 36ef9f79b43df50e25b7f3b68e5c9f6ccf4160e8]. Hugging Face. Accessed 16 November 2026. https://huggingface.co/datasets/McGill-NLP/WebLINX-full