From Full Game State to Live Commentary: A Data-to-Text Pipeline and Corpus for *Age of Empires II*

Siewart van Wingerden[1] **Mariët Theune**[2] **Lorenzo Gatti**[2]
[1]Studio Platypus [2]Human Media Interaction, University of Twente
siewart@studioplatypus.nl {m.theune,l.gatti}@utwente.nl

Abstract

We present SAGE (Semantic Abstraction & Generation from Events), a modular data-to-text pipeline that converts continuous *Age of Empires II* (AoE2) game-state snapshots and events into live text commentary updates, pairing rule-based content selection with a large language model (LLM) realizer. Its content selection operationalizes how human commentators forage the game for information through two independently ablatable strategies: *Expert Insights* (strategic context) and *Event Structuring* (grouping parallel events). Alongside the pipeline we release its underlying corpus: continuous full game-state snapshots for 3,100 ladder and 176 tournament AoE2 matches (142 with paired professional commentary video), to our knowledge the first full-state corpus for a complex esports title. We also report on a preliminary user study with 33 AoE2 viewers.

1 Introduction

Watching esports (competitively played video games) is a mainstream spectator activity, and much of its appeal comes from commentators who narrate, contextualize, and dramatize a match in real time (Kempe-Cook et al., 2019; Li et al., 2020). High-quality live commentary is resource-intensive and concentrated at top tournaments, leaving most competitive matches (and players reviewing them) without narrative guidance. Because the game engine mediates every player action, the full match state is available as data, making automated commentary an attractive data-to-text (D2T) application.

We address three gaps. First, the human elements are often overlooked: few systems model the strategies commentators use to select relevant state for their audience, and engagement is rarely validated with the target audience; instead, reference-based overlap metrics are used. The rare human evaluation is of limited scope and, while raters understand the game, esports viewership is never reported (Ishigaki et al., 2023; Wang and Yoshinaga, 2024; Olarewaju et al., 2020). Second, most esports-commentary systems consume sparse event logs or limited metadata, or target games with a limited state space. For instance, Wang and Yoshinaga (2024) generate commentary from ten high-level event types. Third, datasets available for NLG research generally also focus on these sparse logs or games with limited state complexity, and rarely include paired commentary video; each covers one of these aspects rather than combining them (Wang and Yoshinaga, 2024; Ishigaki et al., 2021; Zhang et al., 2024).

We address all three: SAGE turns continuous AoE2 game-state snapshots into minute-by-minute text updates (online sports commentary) (Pažanin and Petrović, 2022; Lewandowski, 2012). We contribute *a hybrid D2T pipeline* that abstracts a dense, parallel event stream into reportable events, enriches them from state snapshots and a knowledge base, and realizes text with an LLM, using two independently ablatable engagement modules based on commentator strategies (Section 3); the *game-state dataset* consisting of 176 tournament matches, 142 with paired professional commentary video and an extra 3,100 recent ladder matches (Section 4); and *a preliminary user study* ($N = 33$ target-audience spectators) testing each module's effect on engagement and linguistic quality (Section 5).

2 Related Work

Esports and game commentary Early esports-focused systems generated narratives from game logs (e.g., Bardic (Barot et al., 2017)) or grammar-based recaps trained on esports commentator transcripts (Olarewaju et al., 2020). Closest to our setting, Wang and Yoshinaga (2021, 2024, 2025) generate commentary from event records with neural and LLM realizers, but report reduced control over content selection and only limited viewer evaluation. Zhang et al. (2022) compare rule-based and

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 65–73
October 17-21, 2026 ©2026 Association for Computational Linguistics

LLM realization; Zhang et al. (2024) pair events with commentator audio and chat in a multimodal dataset; and Kim et al. (2025) compress match data into LLM-friendly summaries. All of this work targets DotA2 or League of Legends. These are Multiplayer Online Battle Arena (MOBA) games where each player controls one hero and the state and event spaces are far smaller than AoE2's. One genre exception is racing games, which pair video with limited state data (Ishigaki et al., 2021), but again with a narrow event space. For other sports, data-to-text covers simulated (Voelz et al., 1998) and real (Taniguchi et al., 2019) soccer and post-hoc match reports (van der Lee et al., 2017).

Data-to-text pipelines Modular pipelines are often compared with (statistical) end-to-end systems as trading fluency for control (Gatt and Krahmer, 2018). Castro Ferreira et al. (2019) found that a pipeline produced better text than end-to-end generation. Recently, LLMs and new deployment methods such as multi-agent end-to-end systems score better on automatic metrics and semantic adequacy (Osuji et al., 2024, 2025). Surveying neural D2T, Lin et al. (2024) list limited controllability among remaining challenges and conclude that "introducing a content planning stage before text generation can effectively enhance controllability and faithfulness". For game commentary, Zhang et al. (2022) find that "fusing" a rule-based generator with a neural one best balances accuracy and fluency.

Corpora Public esports resources most often cover sparse event logs or compact-state games: League of Legends replay packets (Zhu, 2025), and player trajectories in ESTA (Xenopoulos and Silva, 2022) for a first-person shooter, in which each player controls a single character. Within the same genre as AoE2 we find the SC2EGSet for StarCraft II (Białecki et al., 2023) and AoE2 event logs (Liss et al., 2024). Most are not aimed at NLG or need heavy pre-processing, and several are hard to obtain: the multimodal Game-MUG (Zhang et al., 2024) is event-level and not openly released, and Wang and Yoshinaga (2024) ship tools to regenerate their data rather than the data itself. Ishigaki et al. (2021) provide a racing-game commentary dataset pairing video with structured telemetry, but limited to Japanese and a narrow state space.



Figure 1: *Age of Empires II* as shown in *CaptureAge* during a King of the Desert IV match. CaptureAge creates a live re-rendering from game-state data and low-level events, which SAGE turns into commentary.

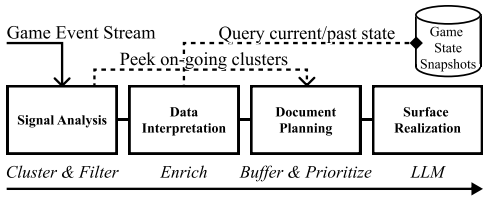


Figure 2: The SAGE pipeline parses a symbolic event stream and game-state snapshots into realized text.

3 The SAGE Pipeline

Age of Empires II (Figure 1) is a real-time strategy game in which opposing players, each representing one of many *civilizations* with different bonuses, simultaneously gather resources, grow an economy of worker units (*villagers*), advance through four *Ages* that unlock stronger units and technologies, and build armies to disrupt each other's economy and claim victory. The map is procedurally generated and partly hidden by a *fog of war*, so players must scout and adapt. Our pipeline must therefore handle economic and military actions unfolding in parallel.

3.1 Architecture

SAGE follows the four-stage architecture of Reiter (2007) (Figure 2). Data arrives incrementally and the system plans and realizes without knowledge of future events, so it runs as if in a live setting, and could be integrated into a live application.

Signal analysis The raw input is a symbolic event stream and periodic full world-state snapshots. An *event* here is emitted from a discrete call the game engine makes while applying its own rules and executing player commands (e.g., a unit taking damage or dying, or a building or technology starting and completing). SAGE pairs them with snapshots windowed at 60 s. We preprocess events in two ways. Damage and death events are clustered into battles

with ST-DBSCAN (Birant and Kut, 2007), a density-based clustering method with a temporal distance. It attributes two events to the same battle when they lie within ε map tiles *and* τ seconds of one another and their neighborhood holds at least `minPts` events ($\varepsilon = 14$ tiles, $\tau = 60$ s, `minPts` $= 10$). This method supports a dynamic number of simultaneous battles, which range from a small skirmish to a full engagement along shifting frontlines. Because the input arrives in chronological order, a cluster can be closed and released once τ seconds pass, enabling live reporting of resolved or stalled battles. Downstream stages can also read an open cluster, or force it closed, to report an ongoing fight. Other event types, such as constructions (Castles, Town Centers, Towers), are kept by filtering on building type and location.

Data interpretation Clustered events are enriched from the snapshot current at event time (or an earlier cached one), simulating live operation. Enrichment follows the Actuator–Environment–Performance+Sensors (A–E–P+S) scheme (Russell and Norvig, 1995; Dodge et al., 2018), reflecting how commentators forage for information (Dodge et al., 2018; Penney et al., 2021; van Wingerden, 2026): the *Actuator* is the event; the *Environment* is the state it occurs in (location, battle participants); *Performance* is who leads on key metrics (e.g., economy); and *Sensors* capture what each side can currently see. To help readers visualize the game map, we add narrative location anchors (names for regions and landmarks) using cardinal and relative directions, so an event at numerical map coordinates is reported at "Blue's forward gold mine".

Enrichment adds and updates the pipeline's data, and text is generated only during realization; however, in the following example we show the steps as text. We start from the Actuator, *"Red's three scouts kill a Villager"*, derived from a battle event cluster. Environment and Performance give *"With Forging now in, Red's three scouts kill a Villager. Red is now three Villagers ahead"*, adding information from the game state snapshots; the location anchors and the game clock give *"16:32 — … on the forward woodline"*.

Planning and realization The planner opens with two introductions: the players and civilizations, and the game map. It buffers enriched events and flushes them when a pending battle's cooldown expires, or at game end. Each content unit is realized by Gemini 2.5 Flash, chosen for low first-token latency, at temperature 1 with a fixed seed. All previously realized text is given as prior turns so references and tone are consistent. Its prompt has three parts: a fixed system instruction, a communicative goal, and the preprocessed JSON with numerical identifiers replaced by names and human-friendly timestamps. Writing style instructions (simple present, copula deletion, spatiotemporal anchoring (Lewandowski, 2012), matching the live-update register) are added to ensure consistency. A module's prompts are added only when that module is enabled, so the final combined prompt changes with the module configuration. The state selected in the previous stages and additional state facts are added for grounding, and we discourage unsupported claims through negative prompting. See Appendix A for an overview.

Realization, with the register effects underlined, turns the data of our earlier example into *"16:32 — With Forging now in, Red's three scouts waste no time — they spot a Villager on the forward woodline! Red surges ahead with a three-Villager advantage."*

3.2 Engagement Modules

Engagement strategies are ablatable modules that add steps to any pipeline stage and can be enabled independently. The two strategies below, and the event and context types the pipeline surfaces, were informed by two preparatory studies (van Wingerden, 2026). In the first, building on analyses of a game in the same genre (Dodge et al., 2018; Penney et al., 2021), we coded a professional AoE2 co-cast (two commentators with complementary roles, Kempe-Cook et al., 2019) for both information source (e.g., descriptive game state or expert domain knowledge) and communicative intent, and two further co-casts for intent alone. This informed our use of the A–E–P+S format found by Dodge et al. (2018), but also highlighted the need for information outside the game state. Utterances describing what is happening are mostly derived from observations of the game state (78%, $n = 86$). But reasoning about the state relies more on external knowledge sources: domain knowledge and analytical inference together cover 86% of strategy discussion ($n = 38$) and 77% of strategy prediction ($n = 34$). In the second study, a formative survey, 57 AoE2 viewers watched a manually created video summary and rated how relevant different event types were for understanding the match. *Battles that cost economic units* was the highest-rated content. In open answers viewers said they would also have

liked to see the missing economic consequences and causal links of those events.

- **Expert Insights** adds the domain knowledge and analytical inference not contained in the game state. Strategic civilization information comes from the AoE2 Fandom wiki and player trivia from Liquipedia,[1] compiled offline into a static knowledge base. A fuzzy-logic detector scores unit and building counts and positions, technology state and worker distribution to label high-level strategies (e.g., "Tower Rush", "Fast Castle"), which are added during data interpretation. Its rule templates were drafted by an LLM and manually verified. Additional facts are injected into the realizer prompt with causal explanations (e.g., "Exposed gold mines are vulnerable to Archers") so it can place events in context.
- **Event Structuring** focuses on the game's parallelism: unlike traditional sports built around a single ball, several fights and economic decisions happen at the same time. It groups related events during document planning so the realization step can express causal connections. The planner buffers events for at most 180 game seconds or five events, and groups those overlapping in time. An LLM system prompt is used to select the most important fights and their causal relations. A separate prompt instruction encourages information asymmetry, which Cheung and Huang (2011) identify as a driver of engagement in StarCraft shoutcasting: the realizer contrasts what each side can see with the all-knowing viewer's perspective to create suspense.

4 The Corpus

We release an extended version of the corpus used in the pipeline, to our knowledge the first full game-state set aimed at NLG for a complex esports title. It comprises the *May 2026* set (3,100 anonymized ranked one-versus-one (1v1) ladder matches from the top 1,000 players by matchmaking rating) and the *King of the Desert IV* (KotD IV) set (176 tournament matches, 142 with commentary video), totaling roughly 120,000 minutes of play and 16 M issued player commands (Table B1). Each match is exported at both 1 s and 60 s snapshot resolution as columnar tables (Table B2); Appendix B describes the contents. The windowed tables track every *entity*, meaning every unit and building on the map, alongside each player's economy, technology and visibility. The events we export from the engine were selected based on the preferences reported in the preparatory viewer survey. The state was extracted with CaptureAge, a match-analysis tool that reconstructs a full match and its advanced statistics from a replay or a running game instance. This corpus enables research on large state spaces, event streams and data frames describing a complex environment in which frequent commands are issued. For 142 matches it pairs these with professional commentary video narrating the most important events, and with the CaptureAge replay for visual inspection. A full snapshot exceeds what a language model can ingest in real time, so a system must select when and what data to use.

5 Preliminary User Study

To investigate the relative contributions of *Expert Insights* and *Event Structuring* to viewer engagement and linguistic quality, we performed a preliminary within-subjects user study using pre-generated texts in the live-update register.

Stimuli and design We generated static commentary using all four configurations (neither module, *Expert Insights*, *Event Structuring*, both) for four KotD IV matches, giving 16 texts of roughly 500–800 words. The texts were used as generated, so they still contain realization artifacts such as repetitions, tense slips and occasional attribution errors. In a Latin-square-style design each participant read four texts, seeing every match and configuration once in randomized order, for 132 evaluated texts.

Participants Our participants are active spectators of competitive AoE2 (32 of 33 watch at least weekly), with a well-above-average self-reported skill rating when playing ($M = 1600$ Elo, $n = 30$), recruited on the AoE2 subreddit in August 2025.

Measures and analysis After each text, participants rated seven engageability motivations (adapted from MSES (Qian et al., 2020), two items each), five linguistic-quality (Callaway and Lester, 2002) and two comprehension items, all on 7-point scales. Per outcome we fitted a linear mixed model with both modules, their interaction and presentation order as fixed effects and random intercepts for match and participant, with sum contrasts, so each coefficient is a deviation from the grand mean. Scale reliability, the model specification, the power analysis and robustness refits are in Appendix C.

[1] https://ageofempires.fandom.com/, https://liquipedia.net/ageofempires/

	Grand mean		Expert Insights			Event Structuring			Model R^2	
Outcome	M	95% CI	β	95% CI	p	β	95% CI	p	marg.	cond.
Engageability (MSES motivations)										
Competitive Nature	3.20	[2.65, 3.74]	**.20**	[.05, .34]	**.008**	.05	[−.10, .19]	.522	.02	.74
Skill Improvement	2.58	[2.09, 3.08]	**.19**	[.07, .30]	**.002**	−.05	[−.17, .07]	.413	.02	.80
Game Knowledge	3.86	[3.22, 4.51]	.14	[−.03, .32]	.109	.02	[−.16, .19]	.854	.01	.73
Skill Appreciation	2.53	[2.08, 2.99]	.10	[−.01, .22]	.082	−.04	[−.15, .07]	.483	.02	.75
Entertaining Nature	2.88	[2.37, 3.38]	.11	[−.03, .25]	.124	.11	[−.03, .25]	.132	.02	.71
Dramatic Nature	2.65	[2.14, 3.17]	.10	[−.05, .25]	.176	.01	[−.14, .16]	.856	.01	.67
Vicarious Sensation	2.79	[2.30, 3.29]	**.25**	[.10, .40]	**.001**	.10	[−.05, .25]	.206	.03	.67
Linguistic quality and comprehension										
Overall Quality	3.82	[3.41, 4.23]	.08	[−.13, .29]	.444	.08	[−.13, .28]	.461	.02	.42
Word Choice	4.28	[3.90, 4.66]	.01	[−.20, .22]	.905	.15	[−.06, .35]	.166	.03	.38
Readability	4.62	[4.04, 5.20]	−.04	[−.26, .17]	.685	−.10	[−.31, .12]	.371	.01	.43
Logicality	5.61	[5.28, 5.93]	.04	[−.13, .21]	.656	**.22**	[.05, .39]	**.013**	.06	.34
Detail Level	4.41	[3.77, 5.05]	**−.24**	[−.45, −.04]	**.022**	.06	[−.14, .27]	.538	.02	.51
Reading Thoroughness	4.56	[3.77, 5.35]	−.15	[−.38, .08]	.202	.14	[−.09, .38]	.218	.02	.39
Understanding	4.46	[3.91, 5.02]	.16	[−.04, .36]	.114	−.12	[−.32, .07]	.217	.02	.56

Table 1: Study results ($N = 33$, 132 ratings), one linear mixed model per outcome, sum contrasts. The intercept is the grand mean over all four configurations on a 7-point scale and each β the deviation from it when that module is on. R^2 is Nakagawa's marginal (fixed effects alone) and conditional (fixed and random). Bold is $p < .05$. No interaction between modules was significant; *Detail Level* has a centered neutral 1 = far too much to 7 = far too little.

Results Engageability was low in absolute terms (Table 1): grand means ranged from 2.53 to 3.86, well below the neutral midpoint for every motivation but *Game Knowledge* (enjoys using own AoE2 knowledge). Linguistic quality was higher (*Logicality* (logical order of text) 5.61, *Readability* (fluency) 4.62): the texts read as coherent and readable. *Expert Insights* significantly raised the three analytical motivations, *Competitive Nature* (enjoys competition, $\beta = .20$, $p = .008$), *Skill Improvement* (reading improves skill, $\beta = .19$, $p = .002$) and *Vicarious Sensation* (as if playing themselves, $\beta = .25$, $p = .001$), and moved perceived *Detail Level* toward "just right" ($\beta = -.24$, $p = .022$). *Event Structuring* improved only *Logicality* ($\beta = .22$, $p = .013$). No module interaction was significant, so the modules act independently. However, main effect sizes were much smaller than participant variance (Appendix C). In closing questions, 16 of 33 believed all four texts were machine-generated ($M = 3.3/4$, $SD = 0.77$).

Open feedback Feedback (96 of 132 evaluations) favored strategic importance and high-level context, and preferred a dry, analytical tone, flagging the live-update register as noise, likely because the updates were presented as a single offline text. Trust was limited: correctly realized events were seen as hallucinated. A game-specific example is a mention of Aztec cavalry, which is normally unobtainable. The system failed to mention the unusual way the player had obtained it, leading a participant to see it as a hallucination. We checked every claim participants flagged as implausible against CaptureAge and the source data; each was a real event.

6 Discussion and Conclusion

Even as language models improve, turning dense full game state into live commentary remains an unresolved problem. We focused on a vertical slice: completing a full pipeline from game state to live commentary, designing from a human analysis and ending with a human evaluation. Using strategies similar to those of human commentators (Dodge et al., 2018; Penney et al., 2021), we saw modest but significant gains in engagement and linguistic quality with the real target audience, yet the baseline scores for engagement are low. Strategic context raised the motivations related to understanding the events, and structuring events improved only perceived order. However, AoE2 viewers engaging with the static text used in our study were looking for information, not "hype". A first direction is to explore the generation of commentary supported by video, rather than text updates. Better ways to express causal links between events, and the context needed to understand them, are another direction indicated by the target audience. Both directions can be explored with the corpus we release alongside the pipeline.

Acknowledgments

We thank CaptureAge for the technology, MembTV and DonDinardoni for the KotD IV matches, and NOVA for the commentary videos.

Limitations

Our stimulus was static text read after the match, a modality that does not exist for AoE2 and is thus unfamiliar to viewers, who are instead used to live commentary supported by video. We designed our system with the goal of such live commentary in mind. The additional 142 paired videos are our route toward this real-time narration. The *Event Structuring* implementation groups events by simple heuristics, leaving room for more sophisticated approaches. We evaluated a single opaque proprietary realizer, limiting both reproducibility and our understanding of the scope of its baseline AoE2 knowledge.

Ethics Statement

The user study was approved by the University of Twente EEMCS ethics committee. Participants gave informed consent and could withdraw their data. The professional commentary videos in the corpus are redistributed with the commentators' and tournament organizers' written permission. The *May 2026* ladder set is anonymized, with player-identifying information removed (including: in-game name, chat messages, exact skill rating, match IDs, lobby titles and profile IDs). Tournament participants are public competitors and the matches are widely available as broadcast recordings. We release the dataset to the public domain (CC0), except fields derived directly from game content (game-statistics data and in-game strings), which remain subject to XBOX's terms, as do the CaptureAge and AoE2:DE recordings, in accordance with XBOX's Game Content Usage Rules.

Supplementary Materials Availability Statement

The SAGE pipeline source, the full corpus, the study stimuli, and the analysis scripts used to produce Table 1 and additional analyses are available at `https://huggingface.co/datasets/Siewart/sage` (DOI `10.57967/hf/10202`) for the corpus and `https://github.com/Siewart/sage` for the source code and study materials.

References

Camille Barot, Michael Branon, Rogelio Cardona-Rivera, Markus Eger, Michelle Glatz, Nancy Green, James Mattice, Colin Potts, Justus Robertson, Makiko Shukonobe, Laura Tateosian, Brandon Thorne, and R. Young. 2017. Bardic: Generating multimedia narratives for game logs. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(2):154–161.

Andrzej Białecki, Natalia Jakubowska, Paweł Dobrowolski, Piotr Białecki, Leszek Krupiński, Andrzej Szczap, Robert Białecki, and Jan Gajewski. 2023. SC2EGSet: StarCraft II esport replay and game-state dataset. *Scientific Data*, 10(1):600.

Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221.

Charles B. Callaway and James C. Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

Gifford Cheung and Jeff Huang. 2011. StarCraft from the stands: Understanding the game spectator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 763–772.

Jonathan Dodge, Sean Penney, Claudia Hilderbrand, Andrew Anderson, and Margaret Burnett. 2018. How the experts do it: Assessing and explaining agent behaviors in real-time strategy games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2023. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, Prague, Czechia. Association for Computational Linguistics.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Lucas Kempe-Cook, Stephen Tsung-Han Sher, and Norman Makoto Su. 2019. Behind the voices: The practice and challenges of esports casters. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Jooyoung Kim, Wonkyung Lee, and Jungwoon Park. 2025. Structured summarization of League of Legends match data optimized for large language model input. *Applied Sciences*, 15(13):7190.

Marcin Lewandowski. 2012. The language of online sports commentary in a comparative perspective. *Lingua Posnaniensis*, 54(1):65–76.

Lingyuan Li, Jirassaya Uttarapong, Guo Freeman, and Donghee Yvette Wohn. 2020. Spontaneous, yet studious: Esports commentators' live performance and self-presentation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2024. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*.

Lukas Liss, Nico Elbert, Christoph M. Flath, and Wil M. P. van der Aalst. 2024. Object-centric event log for Age of Empires game interactions. Dataset, Zenodo. https://doi.org/10.5281/zenodo.13365584.

Oluseyi Olarewaju, Athanasios Kokkinakis, Simon Demediuk, Justus Robertson, Isabelle Nölle, Sagarika Patra, Daniel Slawson, Alan Pedrassoli Chitayat, Alistair Coates, Ben Kirman, Anders Drachen, Marian Ursu, Florian Block, and Jonathan Hook. 2020. Automatic generation of text for match recaps using esport caster commentaries. In *Computer Science & Information Technology (CS & IT)*, pages 117–131.

Chinonso Cynthia Osuji, Brian Timoney, Mark Andrade, Thiago Castro Ferreira, and Brian Davis. 2025. Are multi-agents the new pipeline architecture for data-to-text systems? In *Proceedings of the 18th International Natural Language Generation Conference*, pages 542–553. Association for Computational Linguistics.

Chinonso Cynthia Osuji, Brian Timoney, Thiago Castro Ferreira, and Brian Davis. 2024. Pipeline neural data-to-text with large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 320–329, Tokyo, Japan. Association for Computational Linguistics.

Marin Pažanin and Ivana Petrović. 2022. Sports live text commentary as a hybrid register. *ST-OPEN*, 3:1–29.

Sean Penney, Jonathan Dodge, Andrew Anderson, Claudia Hilderbrand, Logan Simpson, and Margaret Burnett. 2021. The shoutcasters, the game enthusiasts, and the AI: Foraging for explanations of real-time strategy players. *ACM Transactions on Interactive Intelligent Systems*, 11(1).

Tyreal Yizhou Qian, Jerred Junqi Wang, James Jianhui Zhang, and Laura Zhenqiu Lu. 2020. It is in the game: Dimensions of esports online spectator motivation and development of a scale. *European Sport Management Quarterly*, 20(4):458–479.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104.

Stuart Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7096–7103.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Siewart van Wingerden. 2026. One thousand and one tales: Automated Age of Empires II eSports reports from raw real-time game data. Thesis, University of Twente, Enschede, June.

Dirk Voelz, Elisabeth André, Gerd Herzog, and Thomas Rist. 1998. ROCCO: A RoboCup soccer commentator system. In *RoboCup-98: Robot Soccer World Cup II*, pages 50–60. Springer.

Zihan Wang and Naoki Yoshinaga. 2021. From eSports data to game commentary: Datasets, models, and evaluation metrics. In *Proceedings of the DEIM Forum 2021*, H13-3. Japanese Society for the Promotion of Science.

Zihan Wang and Naoki Yoshinaga. 2024. Commentary generation from data records of multiplayer strategy esports game. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 263–271.

Zihan Wang and Naoki Yoshinaga. 2025. Data-to-text generation for esports game commentary of multiplayer strategy game. *Journal of Natural Language Processing*, 32(2):572–597.

Peter Xenopoulos and Claudio Silva. 2022. ESTA: An esports trajectory and action dataset. *Preprint*, arXiv:2209.09861.

Dawei Zhang, Sixing Wu, Yao Guo, and Xiangqun Chen. 2022. MOBA-E2C: Generating MOBA game commentaries via capturing highlight events from the

meta-data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4545–4556, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhihao Zhang, Feiqi Cao, Yingbin Mo, Yiran Zhang, Josiah Poon, and Caren Han. 2024. Game-MUG: Multimodal oriented game situation understanding and commentary generation dataset. *Preprint*, arXiv:2404.19175.

Henry Zhu. 2025. League of Legends decoded replay packets dataset. Dataset, Hugging Face. https://huggingface.co/datasets/maknee/league-of-legends-decoded-replay-packets.

Appendix A. Prompt Structure

Each realizer call is assembled from three layers of prompt plus the data payload (Table A1). The base layer is identical in every call; the communicative-goal layer is selected by the type of content unit being realized; the module layer is present only for the modules enabled in that ablation configuration, including a negative prompt that is added when *Expert Insights* is disabled to keep the realizer from supplying strategic knowledge of its own. The full text of every prompt is released with the pipeline.[2]

Layer	Content	Present if
Base	Role and constraints	always
Base	Output format	always
Base	Output language	always
Base	Output length	always, variable length
Base	Writing style	always
Goal	Match introduction	match-intro
Goal	Map introduction	map-intro
Goal	Event	per event
Module	*EI* role and constraints	*EI* on; per config variant
Module	*EI* negative prompt	*EI* off
Module	*ES* role and constraints	*ES* on; per goal variant
Data	Enriched JSON	always; variable module-specific data
Data	Previous turns	2nd call onwards

Table A1: Prompt setup per call, with *EI* for *Expert Insights* and *ES* for *Event Structuring*. Base and data layers are the same for all configurations with variables replaced (such as a target word count); the goal layer varies with the content unit; the module layer varies with the ablation configuration.

Appendix B. Corpus Contents

Table B1 breaks down the size of the corpus. Player and civilization matchup counts indicate variety and data biases; the KotD IV set is deliberately confined to a single popular map to reduce variability. Table B2 lists the per-match tables exported, written as Apache Arrow IPC (and Parquet) and indexed by a JSON descriptor. We also include the CaptureAge and *Age of Empires II: Definitive Edition* (AoE2:DE) recordings for reference and validation, per-set unit descriptions, and, for KotD IV, match-to-video timings. Both sets are exported at 1 s and 60 s snapshot resolution.

Metric	KotD IV	May 2026
Matches (total/video)	176 / 142	3,100
Video: solo / co-cast	118 / 24	—
Game min. (tot/vid)	6,919 / 5,548	112,713
Commands issued	1.09 M / 0.88 M	15.1 M
Unique players	30	888
Player matchups	42	2,735
Civs (picked / total)	31 / 35	53 / 53
Civ matchups	121	1,266
Maps	Arabia (100%)	Arabia 70%, +15

Table B1: Corpus statistics for the King of the Desert IV and May 2026 sets. "Solo"/"co-cast" = one/two commentators. *1v1 Random Map* is the standard ladder format. Ranked matchmaking draws from a curated pool of 16, Arabia accounting for 70%. The percentage follows from player preferences (they can ban and prefer maps before matchmaking) and closely represents the overall map selection. The civilization totals differ because the game gained civilizations between the two sets.

Table	Kind	Contents
players	windowed	stockpiles, population, Age
entities	windowed	every unit and building: owner, position, hit points, action
masters	windowed	per-owner statistics of each unit and building type
research	windowed	technologies finished, started, queued
world	windowed	game clock, market rates
fog-of-war	windowed	what each player can see
events	series	the engine's event stream
commands	series	the actions each player issued
map	record	terrain and elevation per tile
settings	record	lobby configuration
strings	dict	lookup for the name IDs

Table B2: Per-match data tables. *Windowed* tables are fixed-interval snapshots; *series* are timestamped streams of instants; *record* and *dict* are static. The tables share window, player, entity and string identifier columns, as in a relational database.

Appendix C. Study Questionnaire and Analysis

Questionnaire Participants first answered background questions; the engageability, quality and

[2] https://github.com/Siewart/sage/blob/INLG/packages/sage-core/src/gemini/prompts.ts

comprehension items were then asked after each of the four texts, and closing questions at the end. Every item is rated on a 7-point scale; answers are anonymized and, with the participants' permission, released with the analysis scripts. The exact wording of every item is listed in the repository.[3] The composition of each construct from its items is defined in the analysis scripts.

Engageability is measured as constructs of two questions per motivation, adapted from the MSES (Qian et al., 2020) and reported as averages in Table 1:

- **Competitive Nature:** The enjoyment of high-level competition and rivalry.
- **Skill Improvement:** Watching to learn new strategies and improve one's own gameplay.
- **Game Knowledge:** Interest driven by a deep understanding of the game's mechanics.
- **Skill Appreciation:** Admiring exceptional skills and unique plays by professionals.
- **Entertaining Nature:** Seeking fun and pleasure from the content.
- **Dramatic Nature:** The appeal of uncertainty, comebacks, and underdog stories.
- **Vicarious Sensation:** The feeling of being "in the game" and experiencing it through the player.

The linguistic-quality (Callaway and Lester, 2002) and comprehension items were asked as follows:

- **Overall Quality:** Overall, how would you rate the quality of the text?
- **Word Choice:** How appropriate were the word choices in the match updates?
- **Readability:** How hard was it to read the match updates? (*extremely difficult* to *extremely easy*)
- **Logicality:** How often did you feel the live updates were presented out of order? (*always* to *never*)
- **Detail Level:** How would you rate the level of detail in the match updates? (*far too much* to *far too little*, so 4 is the target rather than 7)
- **Reading Thoroughness:** How much did you skip or skim in order to get to the parts you found interesting? (reversed in analysis)
- **Understanding:** How well do you understand how the match played out?

Except for *Detail Level*, every scale is oriented so that a higher score is better.

[3] https://github.com/Siewart/sage/blob/INLG/studies/eval/questionnaire.md

Reliability Across the seven two-item motivation composites, internal consistency is high to very high: Cronbach's α between .79 and .93, and the correlation between a motivation's two items from .66 to .87.

Model Per outcome we fitted $\textit{Score} \sim \textit{EI} * \textit{ES} + \textit{Order} + (1|\textit{Match}) + (1|\textit{Participant})$, where *EI* and *ES* are the two engagement modules, *Order* is the position of the text in that participant's randomized order, and *Match* and *Participant* are random intercepts. Sum contrasts code modules as ± 1, so the intercept is the grand mean over all four configurations and each β is the deviation from it when that module is on. Comparisons against the neutral midpoint of the scale are Wilcoxon signed-rank tests with Holm correction.

Variance explained Nakagawa's marginal R^2, the part of the variance the modules and order explain by themselves, is between .008 and .062; the conditional R^2, which adds the random effects, is between .341 and .795. The intraclass correlations agree: a median of .59 of the variance lies between participants versus .01 between matches. *Order* was significant for *Overall Quality* ($\beta = 0.19$, $p = .049$).

Robustness A Monte Carlo power analysis (2,000 simulations, using the variability observed in the formative survey) reached 80% power to detect a 0.5-point effect at $\alpha = .05$ with nine participants, both module effects being within-subject. While linear mixed models (LMMs) are commonly applied to 7-point Likert scales, we also fitted each model as an ordinal mixed model, and again with *Match* as a fixed rather than a random effect. Both alternatives left the LMM conclusions unchanged.