The Tatort Test of Intelligence: Towards Narrative Comprehension as a Benchmark for AI

Stefan Kramer[1], **Lennart Baur**[1], **Lars Reinhardt**[1]

[1]Institute of Computer Science, Johannes Gutenberg University Mainz

Abstract

We propose—somewhat tongue-in-cheek, yet with serious implications—a new test for artificial intelligence: the ability to watch a 90-minute episode of the long-running German crime drama *Tatort*, and to explain every relevant detail. This involves reconstructing the evolving social network of characters, identifying their beliefs, desires, and intentions, and, crucially, determining who committed the crime. We argue that this task integrates narrative understanding, commonsense reasoning, social cognition, and theory of mind—and thus provides a uniquely challenging benchmark for AI.

Introduction

From the earliest days of AI, the question of how to evaluate intelligence has been central. Turing's imitation game remains the most famous proposal, yet it has been criticized for various reasons, e.g., the temporal limitation or rewarding shallow deception rather than genuine understanding. Other tests, such as Winograd schemas, Raven's progressive matrices, or benchmarks like GLUE and BIG-Bench, attempt to capture linguistic, logical, or commonsense capabilities.

However, most such tests remain narrow in scope. They fail to capture the rich, open-ended, and ambiguous challenges faced by humans in everyday cognition. Watching and understanding a crime drama like *Tatort* involves narrative comprehension, emotion, deception, shifting alliances, and social reasoning at scale. This suggests a new frontier for evaluating machine intelligence.

The Tatort Challenge

The proposed challenge is simple to state: an AI system must watch a 90-minute episode of *Tatort*, one of Europe's most enduring television series, and then explain it in detail. Tatort (see Figures 1 and 2) is a German-language crime drama that was first aired in 1970. In the meantime, more than 1200 episodes have been shown, with a current average audience of 8.6 million viewers across Germany, Austria, and Switzerland (see Table 1). In 2008, Tatort started in a radio format with, in the meantime, more than 160 episodes (see Table 2).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Tatort is particularly interesting for at least three reasons:

- It is produced by different broadcasting corporations, with competition for quality and originality. This makes the episodes, to some extent, unpredictable and free from typical crime drama clichés.
- Tatort is not just entertainment, but combines crime drama with an undercurrent of social commentary. So, it relates directly to the current state of the world and how it is perceived by humans.
- Tatort is (obviously) a German-language crime drama: Thus, it is not in a low-resource language, but neither it is in the top ten languages spoken world-wide. This makes it an interesting subject of investigation. Clearly, other crime drama could be used for experiments of this kind.

The explanation of Tatort episodes will involve the following tasks:

- Identifying all characters and their relationships;
- Tracking the evolving social network as alliances and conflicts emerge;
- Inferring beliefs, desires, and intentions (BDI models);
- Reconstructing the intrigue, and ultimately identifying the perpetrator;
- Providing justifications grounded in narrative coherence and commonsense knowledge.

In the following, we will elaborate on the theoretical underpinnings of the idea from the social sciences and the humanities, on the related work, and on the first steps we took to illustrate the idea.

Theoretical Underpinning from the Social Sciences and the Humanities

Social Network Analysis

One distinctive feature of crime drama is the continual reshaping of social relations as the intrigue unfolds: suspects become allies, witnesses turn into adversaries, and hidden connections are revealed between apparently unrelated characters. Social Network Analysis (SNA) (Scott 2017) offers a toolkit for capturing and quantifying these dynamics. By representing characters as nodes and their interactions—dialogue, shared events, conflicts, alliances—as

Figure 1: Tatort logo.

Fact	Tatort
First aired	29 November 1970
Number of seasons	54
Number of episodes	1,252
Running time	85–110 minutes (mostly 90)
Countries	Germany, Austria, Switzerland
Average audience (2025)	∼8.6 million viewers

Table 1: Key facts about the *Tatort* TV series.

edges, one can trace the changing structure of the story. Centrality measures can show which characters are in charge of the storyline or the flow of information. Community detection can show groups of people that are working together or groups that are under investigation. Temporal network analysis can show how the strength and valence of relationships change from scene to scene. In the context of the Tatort Test, SNA thus plays a dual role. First, it provides a framework for annotation: human coders and AI systems alike can anchor their interpretations in explicit, scene-by-scene networks. Second, it functions as a benchmark for machine comprehension: to "understand" an episode means not only to follow the dialogue but also to reconstruct how the relational graph transforms in each step. Critically, SNA makes visible the gap between surface narrative and deeper intrigue, since the guilty party often appears peripheral early on and only later emerges as structurally central once hidden connections are revealed. An AI system that can dynamically build and interpret such networks demonstrates a grasp of narrative structure that goes beyond textual paraphrase, encompassing the social fabric in which motives and crimes are embedded.

Figure 2: Tatort investigator teams (as of 2025).

Fact	Radio Tatort
First aired	January 2008
Number of seasons	17
Number of episodes	$\geq$ 160 episodes
Running time	50–55 minutes
Distribution	Broadcasts, media libraries, podcasts

Table 2: Key facts about the *Radio Tatort* series.

The Structure of the Intrigue

French structuralists emphasized that stories are governed not only by surface events but by deep structures. Roland Barthes (Barthes 1977) distinguished between "functions" that drive the plot forward and "indices" that provide atmosphere, cultural cues, or hints. A typical Tatort episode features this interplay: A crime occurs, investigators follow misleading leads, small details acquire meaning only retrospectively, and the ultimate resolution reorders the world of the narrative.

Building on Vladimir Propp's morphology of the folktale (Propp 1968), structuralists noted that narratives often revolve around recurrent roles—villain, hero, helper, false hero—and their transformations. Tatort reinterprets these roles: The inspectors act as subjects seeking truth, while witnesses and suspects switch between helper and opponent positions, challenging the viewer to keep track of the changing roles.

Algirdas Julien Greimas formalized this view with his actantial model (Greimas 1983), which identifies six fundamental functions (Subject, Object, Sender, Receiver, Helper, Opponent). Crime drama exemplifies this distribution: the same character can be an ally in one scene and a deceiver in the next. Detecting such shifts requires an AI not only to parse dialogue but also to reason about latent intentions and conflicting accounts.

Tzvetan Todorov (Todorov 1977) characterized detective fiction as a "double narrative": the concealed narrative of the crime and the overt narrative of the inquiry. Episodes of Tatort follow this structure. They force viewers to piece together the plot from incomplete and conflicting testimonies. Gérard Genette's classifications of order and duration (Genette 1980) highlight the narrative complexity: flashbacks, unreliable witnesses, and shifts in perspective require permanent reorganization of knowledge.

To pass the Tatort Test, an AI has to deal with keeping track of changing social roles, making sense of contradicting stories, and figuring out the deeper logic of the intrigue. These are exactly the kinds of problems that regular benchmarks do not deal with, but that are important aspects of human intelligence.

The Intentional Stance

Among the most enduring contributions of philosophy to cognitive science is Daniel Dennett's concept of the *intentional stance* (Dennett 1987). The idea is deceptively simple: rather than explaining a system's behavior in terms of its physical make-up (the physical stance) or functional design (the design stance), we often predict and interpret actions by treating the system as if it had beliefs, desires, and intentions. This strategy underlies much of human social cognition. When we watch a character in Tatort linger over a telephone before dialing, we do not merely note the physical motions of hand and receiver. We infer hesitation, inner conflict, perhaps even fear of exposure. Such inferences allow us to anticipate actions ("she will eventually call the accomplice") and to explain outcomes ("he lied because he feared discovery").

For an artificial system, adopting the intentional stance is not trivial. A large language model can recite summaries or generate plausible dialogue, but the deeper challenge is to maintain a coherent attribution of mental states over ninety minutes of shifting alliances, hidden motives, and red herrings. An agent that passes the Tatort Test would need to recognize not only what each character says and does, but also what they *mean*, what they *know* or *fail to know*, and how these mental states evolve as new evidence emerges. This entails bridging perception (from visual and textual cues) with a model of agency, deception, and social expectation.

The Analogical Theory of Mind (ATOM)

The intentional stance provides a perspective for interpretation. However, the question remains how beliefs and desires are actually represented and reasoned about. The Analogical Theory of Mind (ATOM) (Rabkina 2017) suggests that humans understand other agents' mental states by means of analogy with their own cognitive structures and provides a mechanism for it. When we see a detective struggle with a case, we map our own experiences of problem solving and uncertainty onto the fictional character. Similarly, when a suspect provides an alibi, we project our own social knowledge and knowledge of deception to judge its plausibility.

In the Tatort setting, an AI system could use ATOM-like mechanisms to generate and revise hypotheses about characters. For example, when a person insists too vehemently on their innocence, the system might find similarities with patterns of overcompensation in human testimony datasets, leading it to suspect concealed guilt. Such analogical reasoning could go beyond surface pattern recognition. This would require constructing mental models that can be flexibly reinterpreted as the narrative evolves.

Specific attention has been given to crime stories, though mostly in information extraction and forensic linguistics rather than holistic understanding. The intentional stance has influenced cognitive architectures, while ATOM provides a formal bridge between analogy and "mindreading". Yet, no benchmark to date combines all these strands in a single, unified challenge.

Narrative Understanding as a Cognitive Test

The intentional stance and ATOM combined provide a theoretical basis for understanding why comprehending crime drama is a demanding test of intelligence. To pass the Tatort Test, you need to (i) see characters as purposeful agents, (ii) give them beliefs, desires, and intentions, (iii) connect what you see them do to similar things you have seen before, and (iv) put all of these pieces together into a clear story. This challenge tests the kind of reasoning that is dynamic, socially entrenched, and structured like a story, which is different from question-answering standards that only evaluate facts. In this way, it serves as a replacement for, or addition to, current AI tests, while retaining a playful nod to the longstanding fascination with comparing machines to humans.

Narratives, Commonsense, and the Human Condition

Narratives are more than sequences of events: They are mirrors of the human condition, encoding themes of love, betrayal, greed, justice, and redemption. In Tatort, the crime is never only a puzzle to be solved but a social drama embedded in cultural, moral, and emotional contexts. Understanding such stories requires commonsense knowledge about how people live, argue, and reconcile, as well as about institutions such as families, workplaces, and the law. The guilty party is often revealed not only through forensic evidence but also through cues of motivation, resentment, or despair—cues that are legible only against a backdrop of shared human experience. This is a challenge for current AI systems, because they have to go beyond textual or visual recognition to reason about unstated norms: what counts as suspicious behavior, why jealousy leads to violence, or how shame motivates concealment. Such reasoning entails integrating commonsense ontologies with models of social interaction, while also handling ambiguity and contradiction. Crucially, narrative comprehension demands sensitivity to

irony, tragedy, and moral tension, features that frequently resist simple logical formalization. A system that can navigate a Tatort episode and articulate its significance would therefore demonstrate not only linguistic and inferential competence, but also a capacity to grasp the cultural and ethical dimensions that make us human.

Related Work

Research relevant to the Tatort Test spans several domains: theory of mind (ToM) in humans and machines, benchmarks for narrative and crime-story understanding, character modeling, and methodologies for prompting large language models. Here we review these strands and highlight their connections.

Theory of Mind and the Analogical Perspective

The Theory of Mind (ToM) has been a central element of cognitive science and psychology for a long time. Rabkina's *Analogical Theory of Mind* (AToM) (Rabkina 2017) proposes a cognitive architecture that builds upon the Theory of Mind, including a mechanism for perspective-taking and mental state attribution. Beaudoin, Leblanc, and Gagner's examination of ToM measures for young children (Beaudoin et al. 2020) demonstrates the extensive range of tasks employed to assess this skill ("mindreading"), encompassing false-belief tests and more contextually grounded, socially embedded scenarios. Recent work by Ma, Sansom, and Peng takes this into the AI domain (Ma et al. 2023), surveying how large language models exhibit (or fail to exhibit) a situated form of ToM, and calling for holistic evaluations that go beyond toy tasks. These contributions provide theoretical grounding for the Tatort Test, which combines analogical reasoning with situated, socially entangled story contexts.

Narrative and Crime-Story Benchmarks

Several benchmarks have been suggested that share similarities with our idea. The Conan Benchmark (Zhao et al. 2024) focuses on the identification of character relation graphs. The WhoDunIt dataset (Gupta 2025) directly targets crime-story comprehension: it does so with augmentations of classical detective stories, with more than half of them shorter than 50 pages. Chatter and ChatterEval (Baruah and Narayanan 2025) are benchmarks for character attribution. They evaluate how models sustain coherence and track speaker goals across turns. These datasets suggest a growing recognition that complex narratives present unique challenges to AI. However, they remain limited in scope compared to the multi-layered intrigue of full-length televised drama or even compared to its radio version.

Character Modeling and Profiling

To understand a crime drama, one needs to keep track of what happens and also keep track of the characters, their goals, and how their plans change. Getachew and Saparov's (Getachew and Saparov 2025) StorySim framework frames ToM tasks as simulation exercises where agents must predict and explain states in narrative contexts. Such work underscores that character-centered reasoning is important for story understanding, and it is similar to the Tatort Test's emphasis on evolving social networks and belief states.

Scenario-Based Evaluation and Series Benchmarks

Researchers have begun investigating benchmarks with multiple episodes, which goes beyond the task of understanding a single story. SeriesBench (Zhang et al. 2025) evaluates models on multi-episode narrative comprehension, emphasizing long-range dependencies, consistency, and world-building. This move toward interconnected narratives highlights the growing need for benchmarks that capture temporal continuity and shifting contexts. These are also qualities that make Tatort episodes challenging and rewarding as a material for an intelligence test.

Prompting Methodologies

Finally, work on prompting large language models provides methodological guidance. (Sahoo et al. 2025) survey techniques and guidelines for prompt engineering. They emphasize the need for clear task descriptions, explicit constraints, and fixed output formats. In the Tatort Test, such considerations are central: to extract evolving social networks or belief states from a model, instructions must be carefully crafted to elicit structured, comparable outputs that support validation and use in measures such as those of inter-annotator agreement.

Summary

In sum, the Tatort Test draws on several strands of prior work: analogical and situated theories of mind, narrative reasoning benchmarks, character modeling, scenario-based evaluations, and best practices in prompting. Yet it integrates them into a single, unified challenge that is longer, richer, and more entangled than existing tests. By requiring systems to engage with ninety minutes of intrigue, social complexity, and human motivation, it aims to push the study of AI narrative understanding beyond toy settings and toward a closer approximation of genuine human comprehension.

Research on story understanding in AI has a long history, from early narrative reasoning systems to modern large language models tested on children's stories and movie scripts. Commonsense reasoning (e.g., ATOMIC, ConceptNet) provides partial scaffolding, but narrative comprehension remains elusive.

Specific attention has been given to crime stories, though mostly in information extraction and forensic linguistics rather than holistic understanding. The intentional stance has influenced cognitive architectures such as ATOM. Yet, no benchmark to date combines all these strands in a single, unified challenge.

First Steps Towards the Tatort Test

To take the first steps towards the realization of a rudimentary Tatort Test, we decided to focus on the Radio Tatort first. We contacted the broadcasting corporation SWR and obtained four different scripts: "Dillinger muss sterben" with 33 scences on 57 pages, "Nacht der vergessenen Sterne" with 43 scenes on 65 pages, "Teufel komm raus" with 28

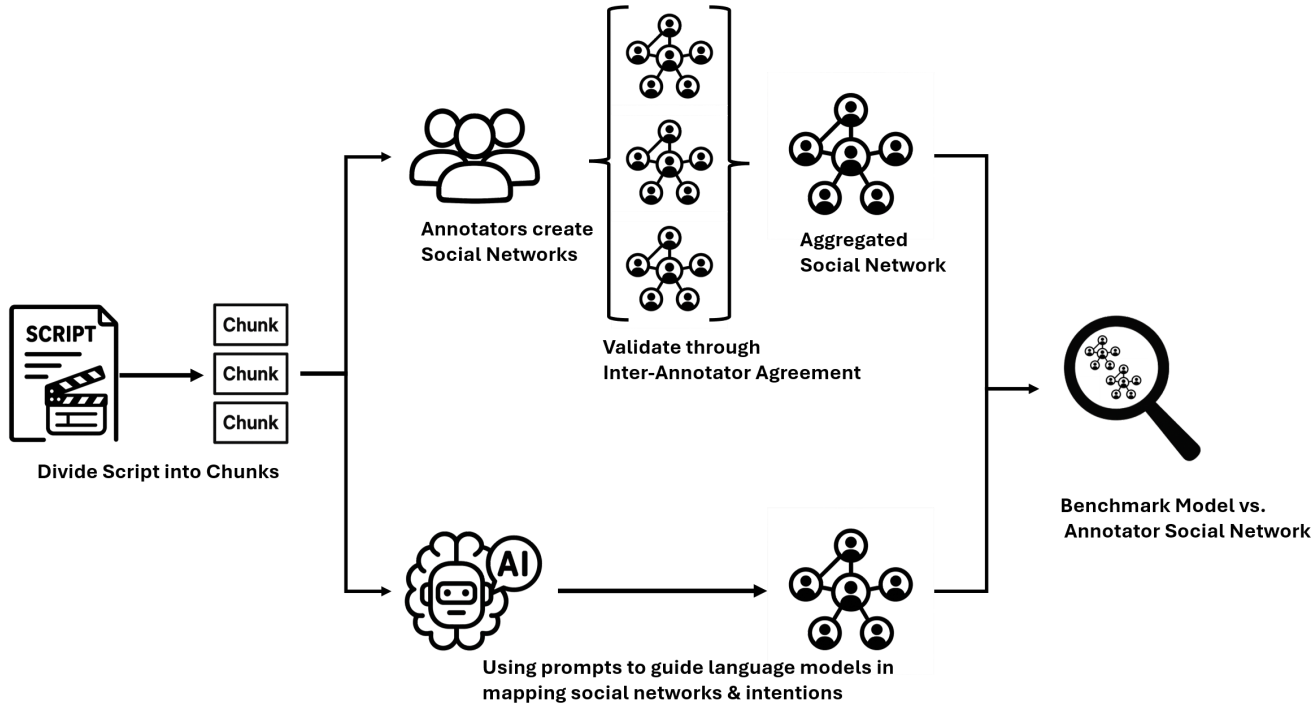
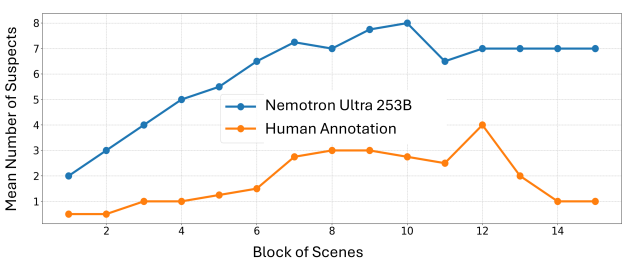
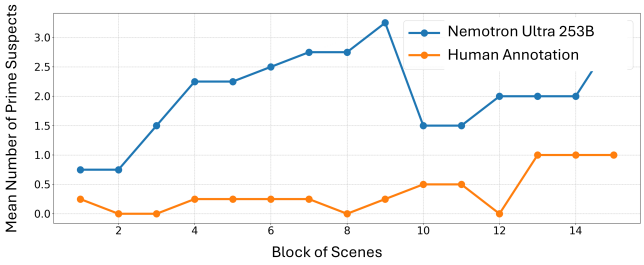


Figure 3: Initial steps towards a Tatort Test for Radio Tatort: workflow for annotations and experiments



(a) Mean number of suspects identified by scene block by humans and by LLM.



(b) Mean number of prime suspects identified by scene block by humans and by LLM.

Figure 4: Comparison humans and a language model in identifying suspects and prime suspects.

scenes on 65 pages, and "Terrorvögel" mit 28 scenes on 68 pages. We followed the following steps (see also Figure 3):

- Perform four independent human annotations of each episode. The pool of annotators consists of seven persons.
- Every three scenes, reconstruct the social network, update BDI states, and record key events. In total, 367 blocks of scenes have been annotated.
- Use language models with carefully designed prompts to aid in reconstructing networks and intentions. In our case, we used the locally hosted Nemotron Ultra 253B to avoid problems with data provenance, where the language model may have read something about the Tatort episode of the test already.
- Validate annotations through inter-annotator agreement.
- Benchmark model performance against this gold standard.

An excerpt of results is shown in Figure 4. As can be seen, the set of suspects and prime suspects identified by the LLM is usually much larger than the sets identified by humans. Many more results from our initial study show the strengths and weaknesses of current LLMs for this task. We intend to publish the current benchmark at a later point in time at https://github.com/kramerlab/.

The approach highlights the dual role of humans and machines: annotators provide ground truth, while LLMs attempt to capture evolving, nuanced narratives. Success hinges on prompt design, narrative decomposition, and robust validation.

Conclusions

The Tatort Test of Intelligence is admittedly playful, yet it highlights a serious research agenda: advancing AI from shallow benchmarks toward genuine narrative and social understanding. As a first step, we annotated four different episodes of the Radio Tatort and tested the ability of current LLMs to follow the narrative. It has to be noted that the full Tatort Test, with televised crime drama, has be complemented with precise performance measures and thresholds to make it a fully specified test for intelligence.

If an AI can pass the Tatort Test, it will have demonstrated not only technical sophistication but also a grasp of stories, society, and ultimately, humanity.

References

Barthes, R. 1977. Introduction to the Structural Analysis of Narratives. In Heath, S., ed., *Image, Music, Text*, 79–124. New York: Hill and Wang. Originally published in *Communications* 8 (1966).

Baruah, S.; and Narayanan, S. 2025. CHATTER: A character-attribution dataset for narrative understanding. In *Proceedings of the The 7th Workshop on Narrative Understanding*, 52–63.

Beaudoin, C.; Leblanc, ; Gagner, C.; and Beauchamp, M. H. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, Volume 10 - 2019.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Genette, G. 1980. *Narrative Discourse: An Essay in Method*. Ithaca, NY: Cornell University Press. Originally published in French as *Discours du récit* (1972).

Getachew, N.; and Saparov, A. 2025. Language Models Might Not Understand You: Evaluating Theory of Mind via Story Prompting. In *Proceedings of the COLM Workshop on Social Simulation with LLMs*.

Greimas, A. J. 1983. *Structural Semantics: An Attempt at a Method*. Lincoln: University of Nebraska Press. Originally published in French as *Sémantique structurale* (1966).

Gupta, K. 2025. WHODUNIT: Evaluation benchmark for culprit detection in mystery stories. ArXiv preprint arXiv:2502.07747.

Ma, Z.; Sansom, J.; Peng, R.; and Chai, J. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1011–1031. Singapore: Association for Computational Linguistics.

Propp, V. 1968. *Morphology of the Folktale*. Austin: University of Texas Press, second edition. Originally published in Russian, 1928.

Rabkina, I. 2017. AToM: An Analogical Theory of Mind. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 5203–5204.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2025. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. ArXiv preprint arXiv:2402.07927.

Scott, J. 2017. *Social Network Analysis*. London: SAGE Publications, fourth edition.

Todorov, T. 1977. *The Poetics of Prose*. Ithaca, NY: Cornell University Press. Originally published in French as *Poétique de la prose* (1971).

Zhang, C.; Lei, Y.; Liu, Z.; Leng, H.; Liu, S.; Gao, T.; Liu, Q.; and Wang, Y. 2025. SeriesBench: A Benchmark for Narrative-Driven Drama Series Understanding. In *2025 Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, 28995–29004.

Zhao, R.; Zhu, Q.; Xu, H.; Li, J.; Zhou, Y.; He, Y.; and Gui, L. 2024. Large Language Models Fall Short: Understanding Complex Relationships in Detective Narratives. In *Findings of the Association for Computational Linguistics: ACL 2024*, 7618–7638. Bangkok, Thailand.