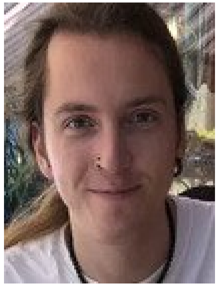
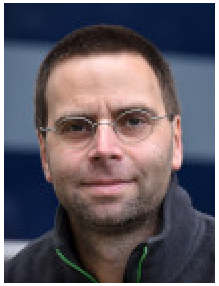
Who Will Evaluate the Evaluators?
Exploring the Gen-IR User Simulation Space

Johannes Kiesel[1]

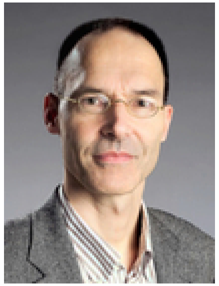


Marcel Gohsen[1]

Nailia Mirzakhmedova[1]



Matthias Hagen[2]



Benno Stein[1]

Bauhaus-Universität Weimar [1]
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA [2]
Webis [1,2]

How to Evaluate Conversational Systems?

A Case for User Simulation

User: Can you explain what the Turing Test is, its importance, and provide some background information on it?

How to Evaluate Conversational Systems?

A Case for User Simulation

User: Can you explain what the Turing Test is, its importance, and provide some background information on it?

System: The Turing Test, introduced in Alan Turing's 1950 paper 'Computing Machinery and Intelligence,' is a benchmark for AI that evaluates a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. It's important as it helps define and measure progress in developing thinking machines, although it has limitations such as subjectivity and the need for clear definitions of 'think' and 'machine'. The Turing Test concept has also been adapted and expanded, like in the Reverse Turing Test.

How to Evaluate Conversational Systems?

A Case for User Simulation

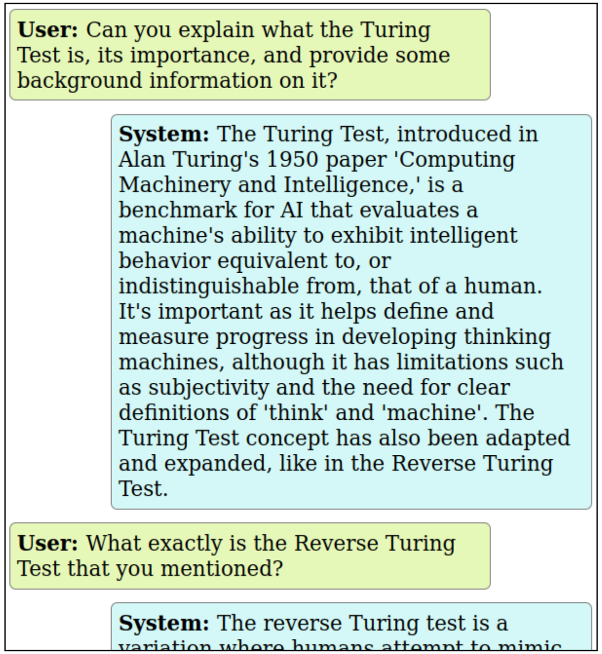
User: Can you explain what the Turing Test is, its importance, and provide some background information on it?

System: The Turing Test, introduced in Alan Turing's 1950 paper 'Computing Machinery and Intelligence,' is a benchmark for AI that evaluates a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. It's important as it helps define and measure progress in developing thinking machines, although it has limitations such as subjectivity and the need for clear definitions of 'think' and 'machine'. The Turing Test concept has also been adapted and expanded, like in the Reverse Turing Test.

User: What exactly is the Reverse Turing Test that you mentioned?

How to Evaluate Conversational Systems?

A Case for User Simulation



User: Can you explain what the Turing Test is, its importance, and provide some background information on it?

System: The Turing Test, introduced in Alan Turing's 1950 paper 'Computing Machinery and Intelligence,' is a benchmark for AI that evaluates a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. It's important as it helps define and measure progress in developing thinking machines, although it has limitations such as subjectivity and the need for clear definitions of 'think' and 'machine'. The Turing Test concept has also been adapted and expanded, like in the Reverse Turing Test.

User: What exactly is the Reverse Turing Test that you mentioned?

System: The reverse Turing test is a variation where humans attempt to mimic

Main problems of human evaluation:

- Not scalable
- Not replicable

Main problems of static test collections:

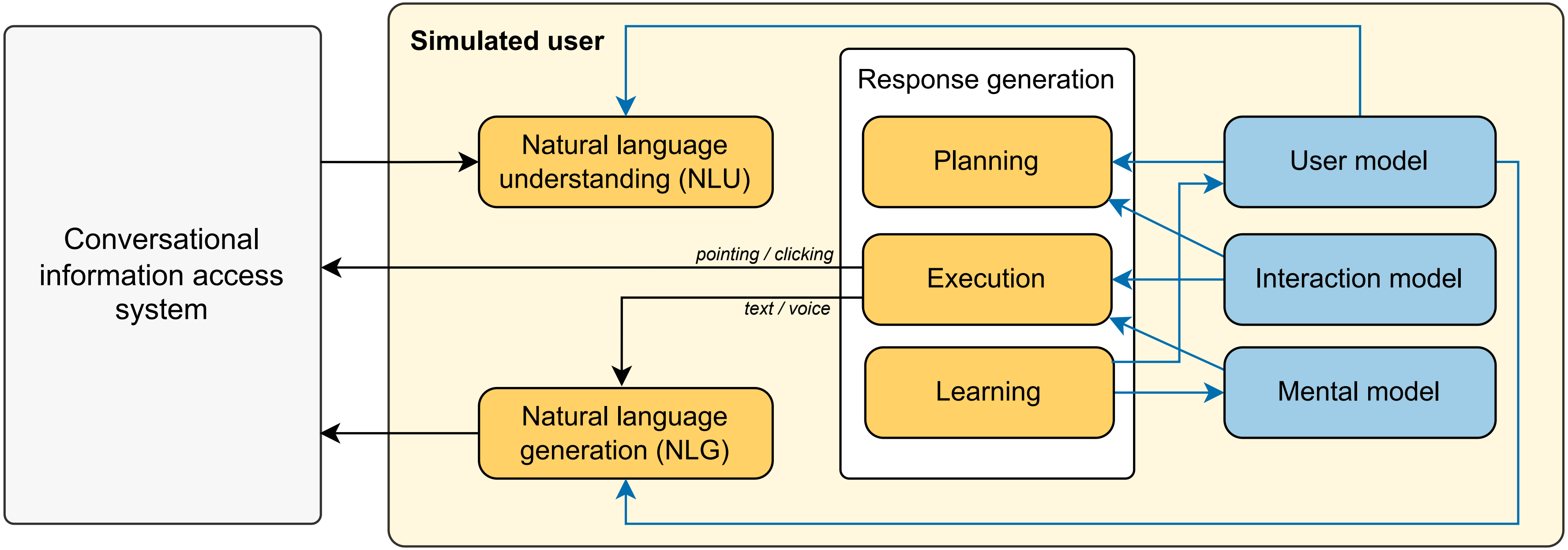
- Low dialog coverage
- No multi-turn interactions

Solution: User simulation

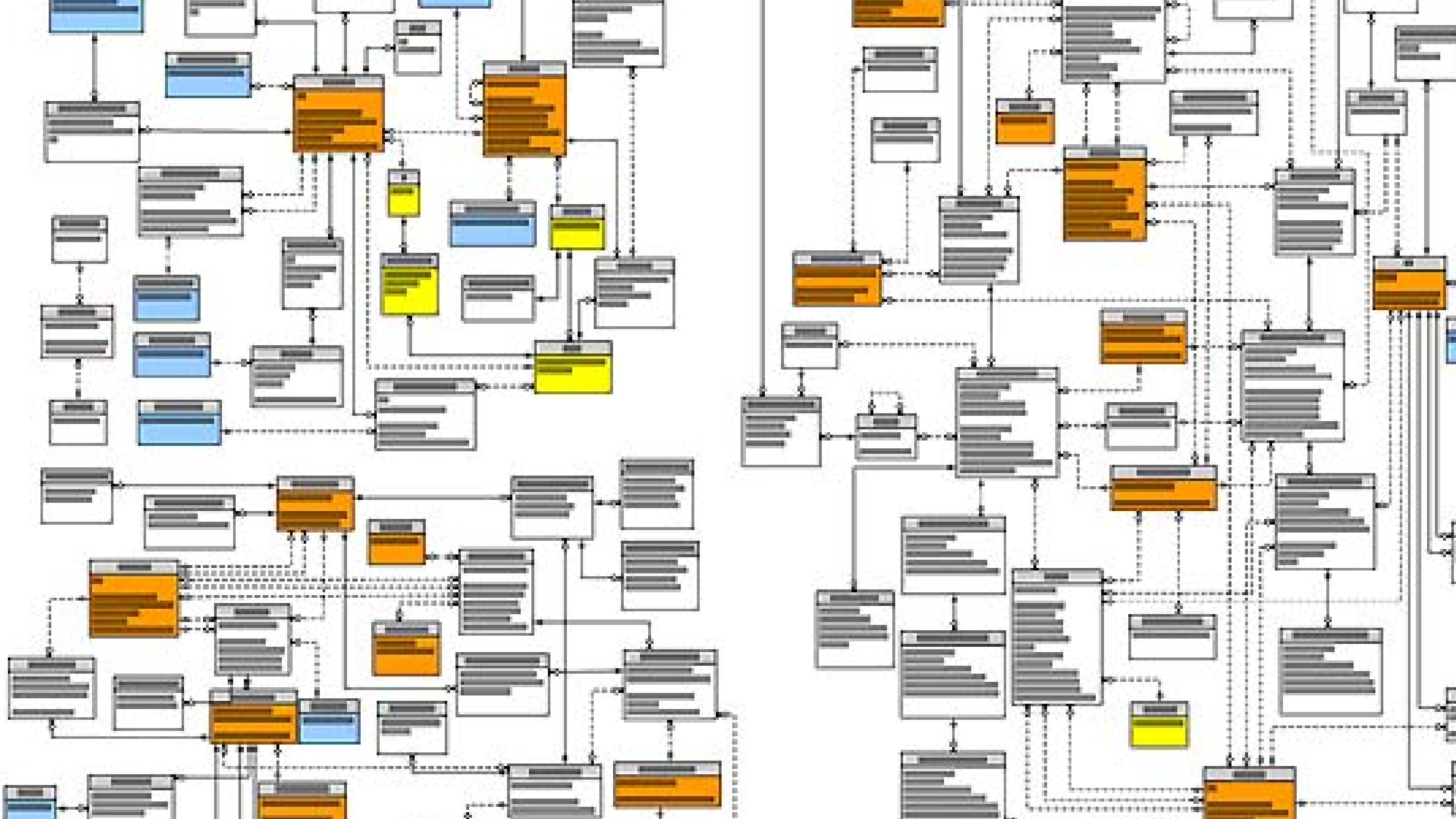
Replace 👤 with 🤖 (including judgement)

How to Evaluate Conversational Systems?

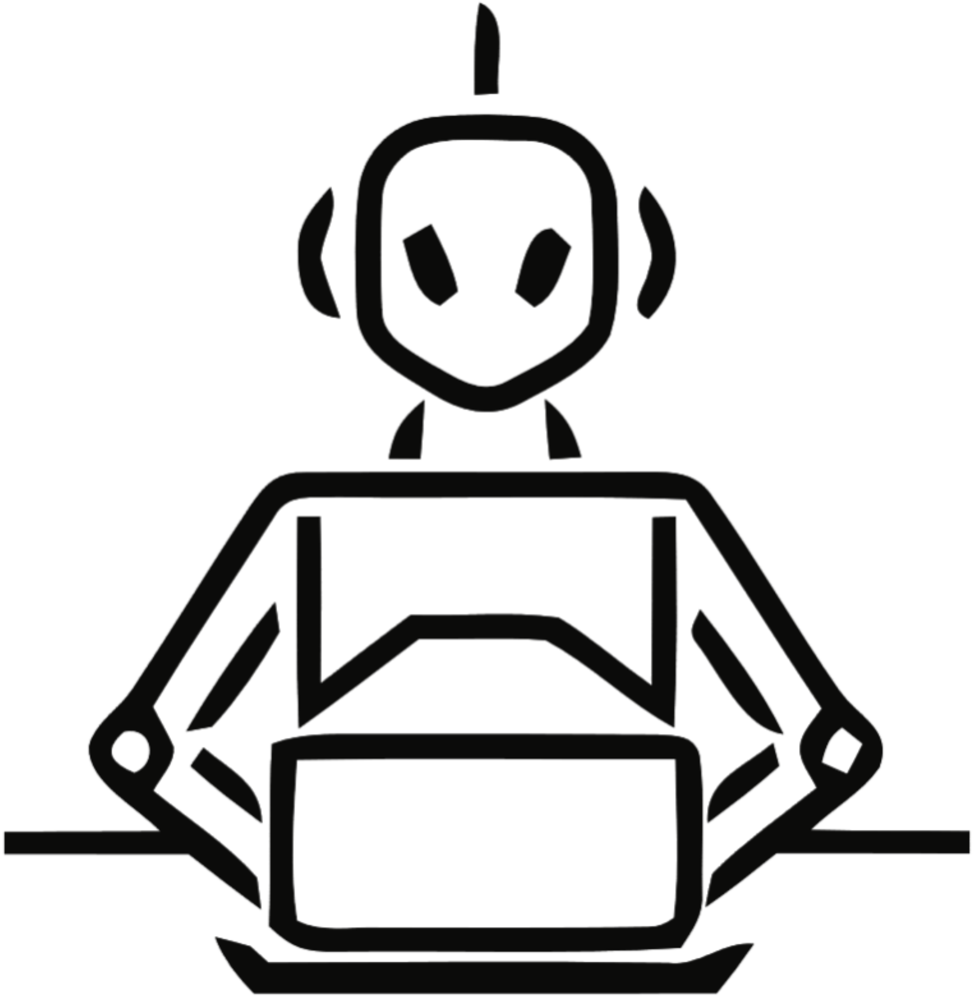
A Case for User Simulation



User simulation architecture proposed by K. Balog (2021; not implemented)



GenIRSim

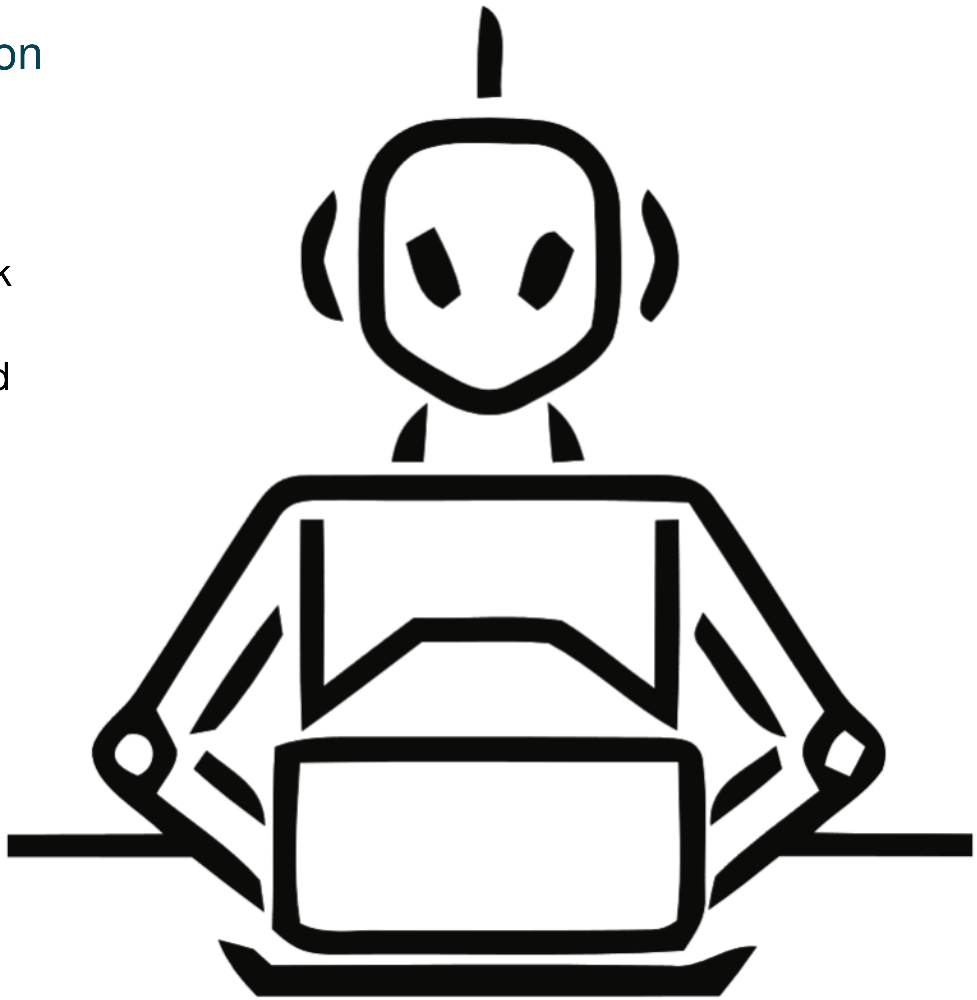


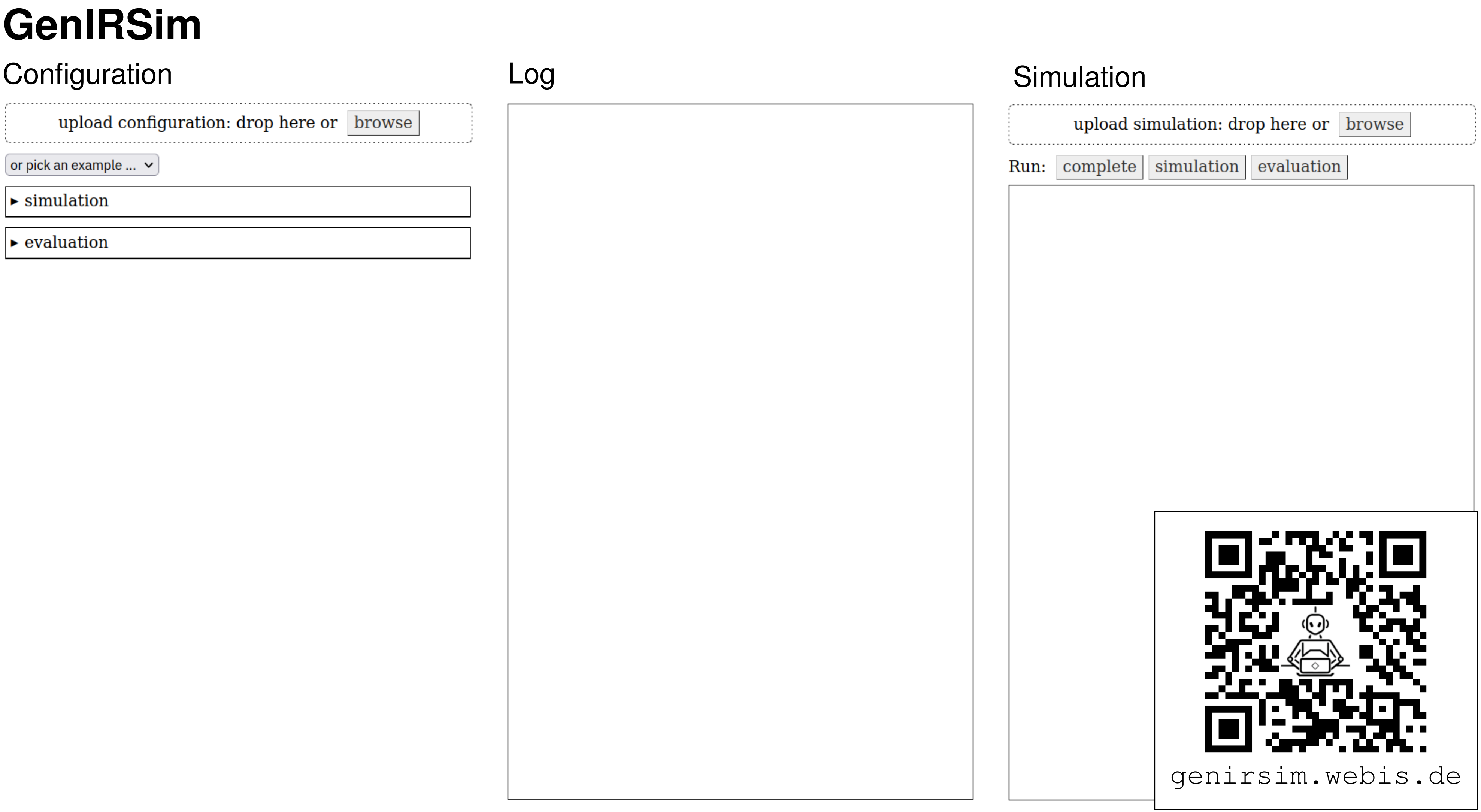
GenIRSim

Flexible simulation and evaluation framework for generative IR

- Command line and web interface
- File-based configuration and quick deployment
- Interlinked simulation, search, and evaluation
- Flexibile and extensibile

genirsim.webis.de





GenIRSim

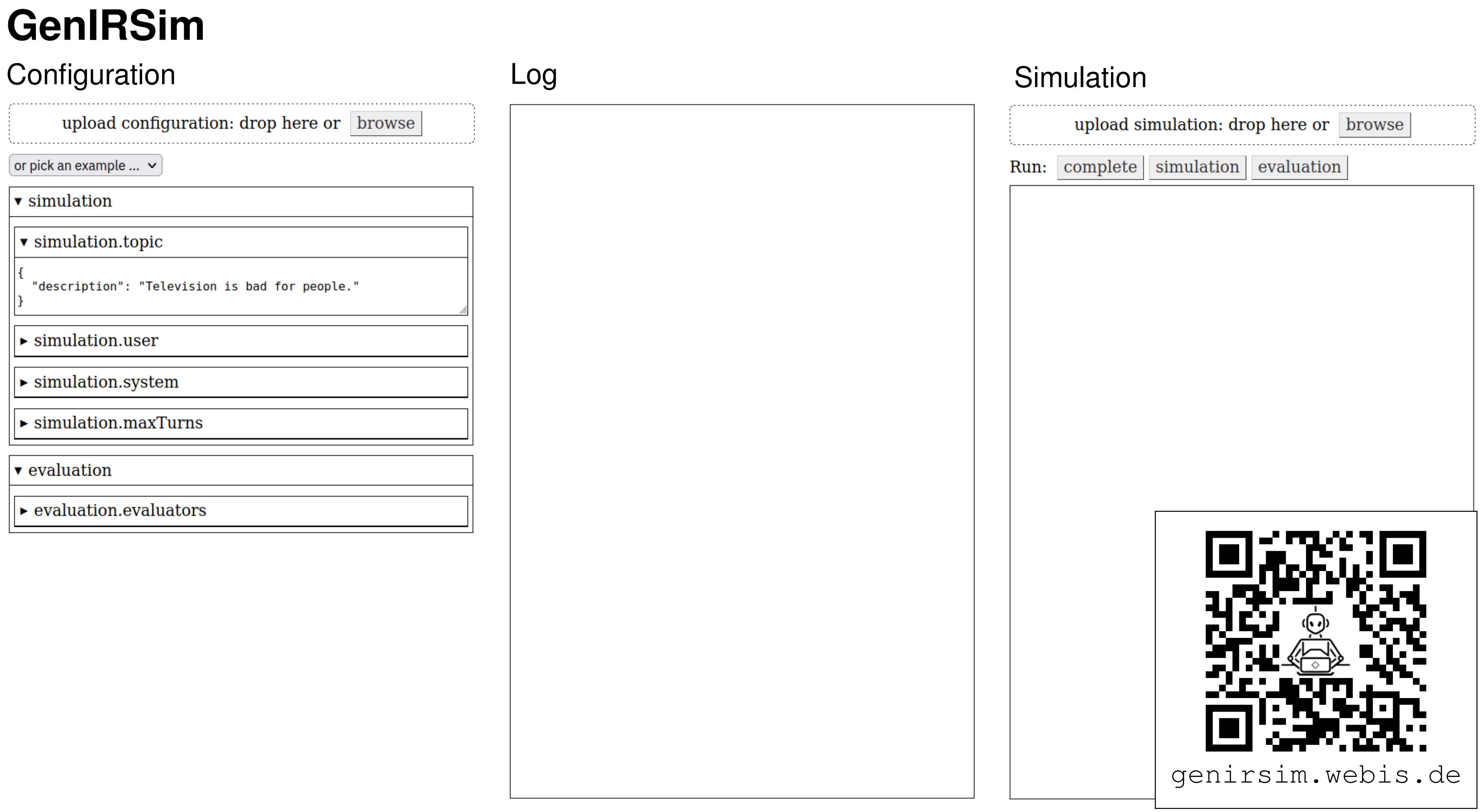
Configuration

Log

Simulation

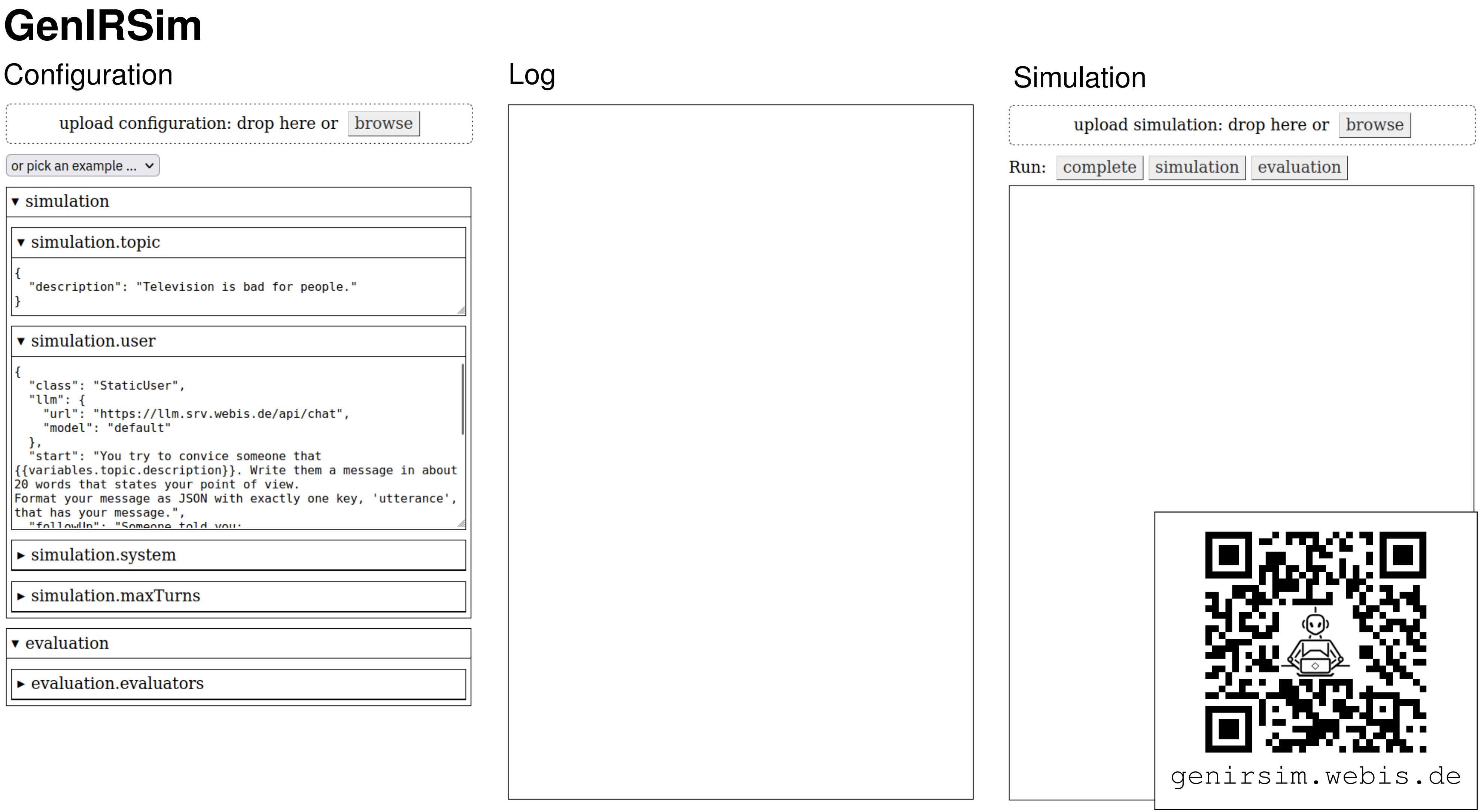
upload configuration: drop here or browse
or pick an example ...
▸ simulation
▸ evaluation
upload simulation: drop here or browse
Run: complete simulation evaluation

genirsim.webis.de



```
{
  "description": "Television is bad for people."
}
```

genirsim.webis.de



GenIRSim

Configuration

▾ simulation

▾ simulation.topic

```
{
  "description": "Television is bad for people."
}
```

▾ simulation.user

```
{
  "class": "StaticUser",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "start": "You try to convice someone that
{{variables.topic.description}}. Write them a message in about
20 words that states your point of view.
Format your message as JSON with exactly one key, 'utterance',
that has your message.",
  "followUp": "Someone told you:
```

▸ simulation.system

▸ simulation.maxTurns

▾ evaluation

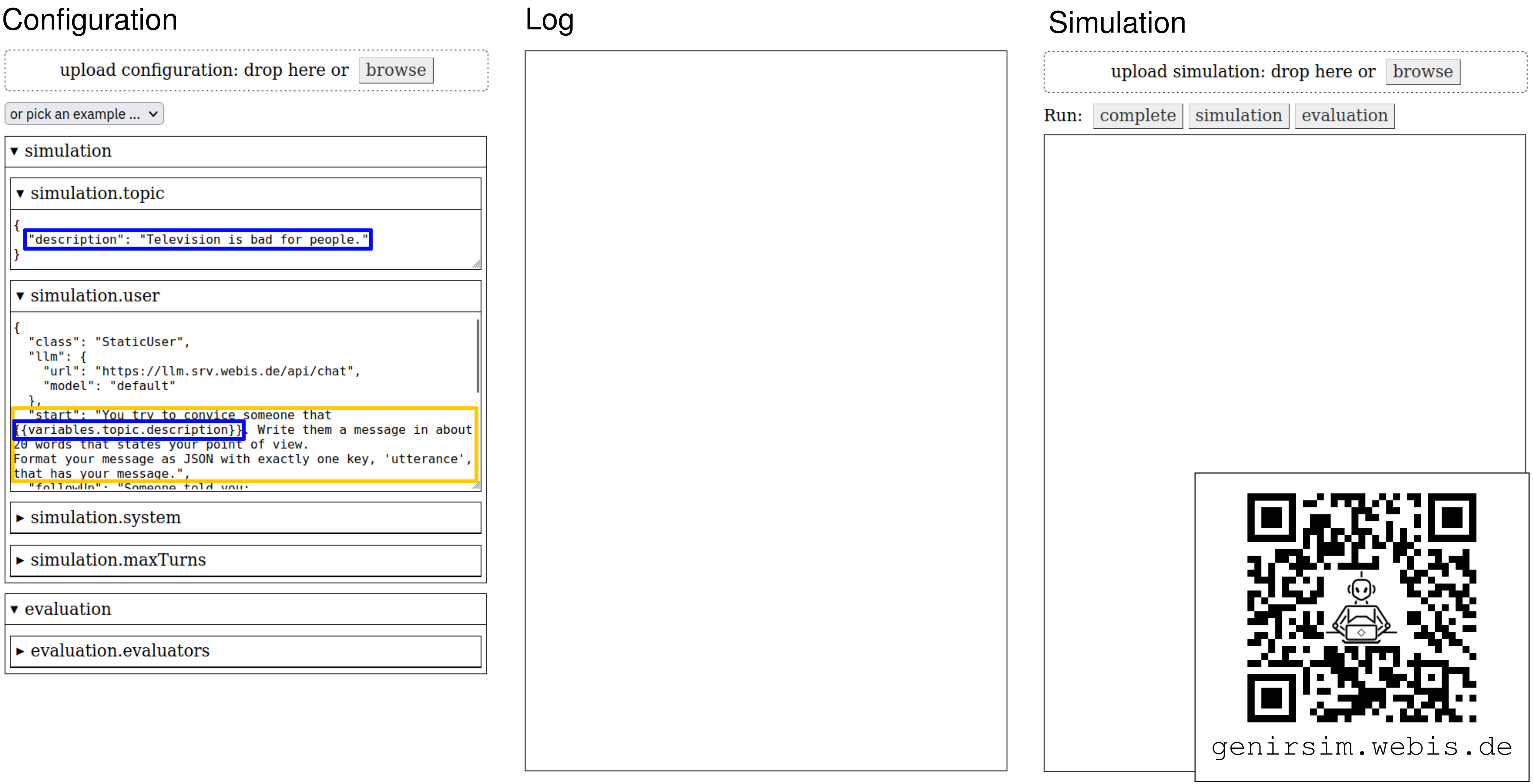
▸ evaluation.evaluators

Log

Simulation

genirsim.webis.de

GenIRSim



simulation.topic

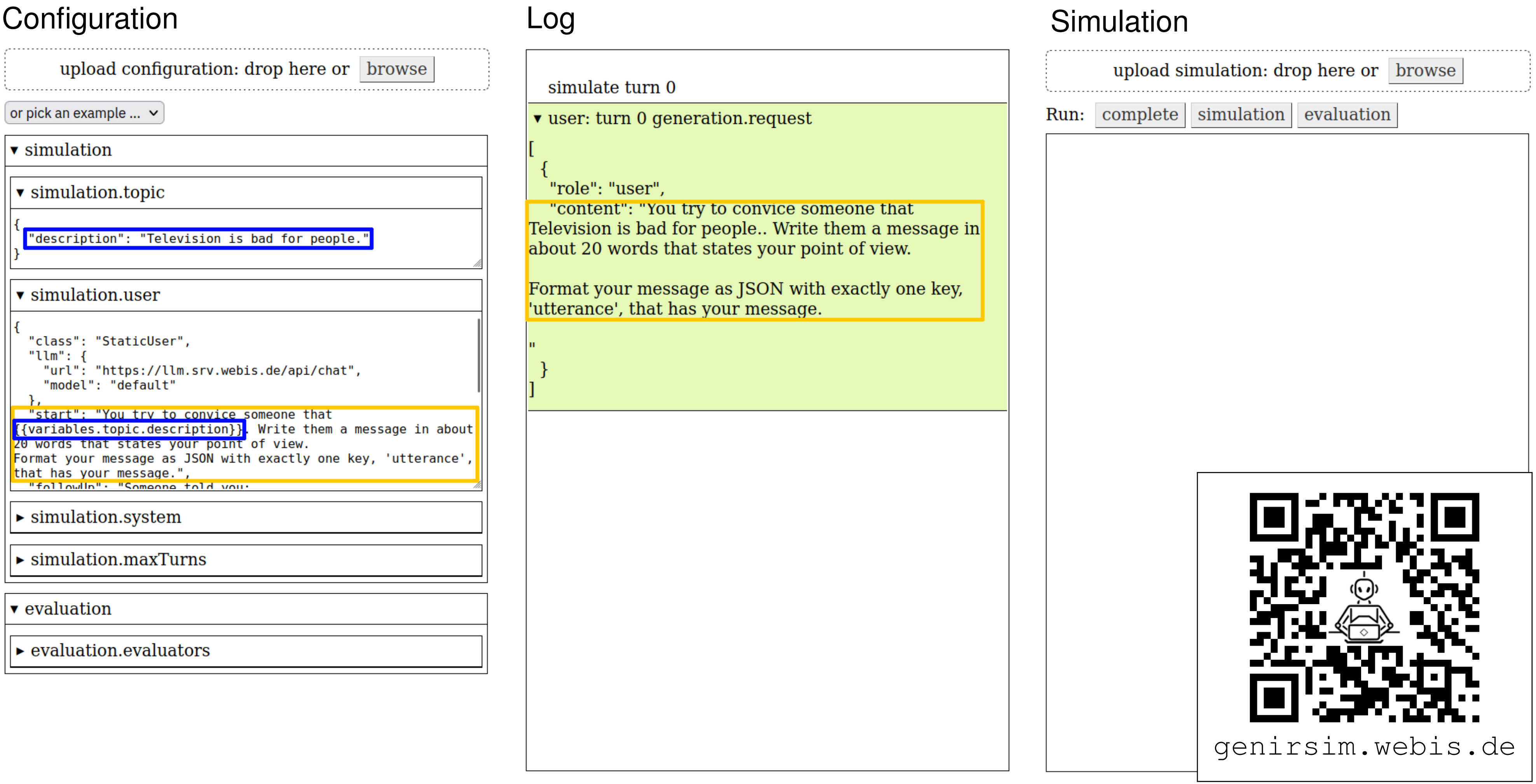
```
{
  "description": "Television is bad for people."
}
```

simulation.user

```
{
  "class": "StaticUser",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "start": "You try to convice someone that
{{variables.topic.description}}. Write them a message in about
20 words that states your point of view.
Format your message as JSON with exactly one key, 'utterance',
that has your message.",
  "followUp": "Someone told you:
```

genirsim.webis.de

GenIRSim



```
{
  "description": "Television is bad for people."
}
```

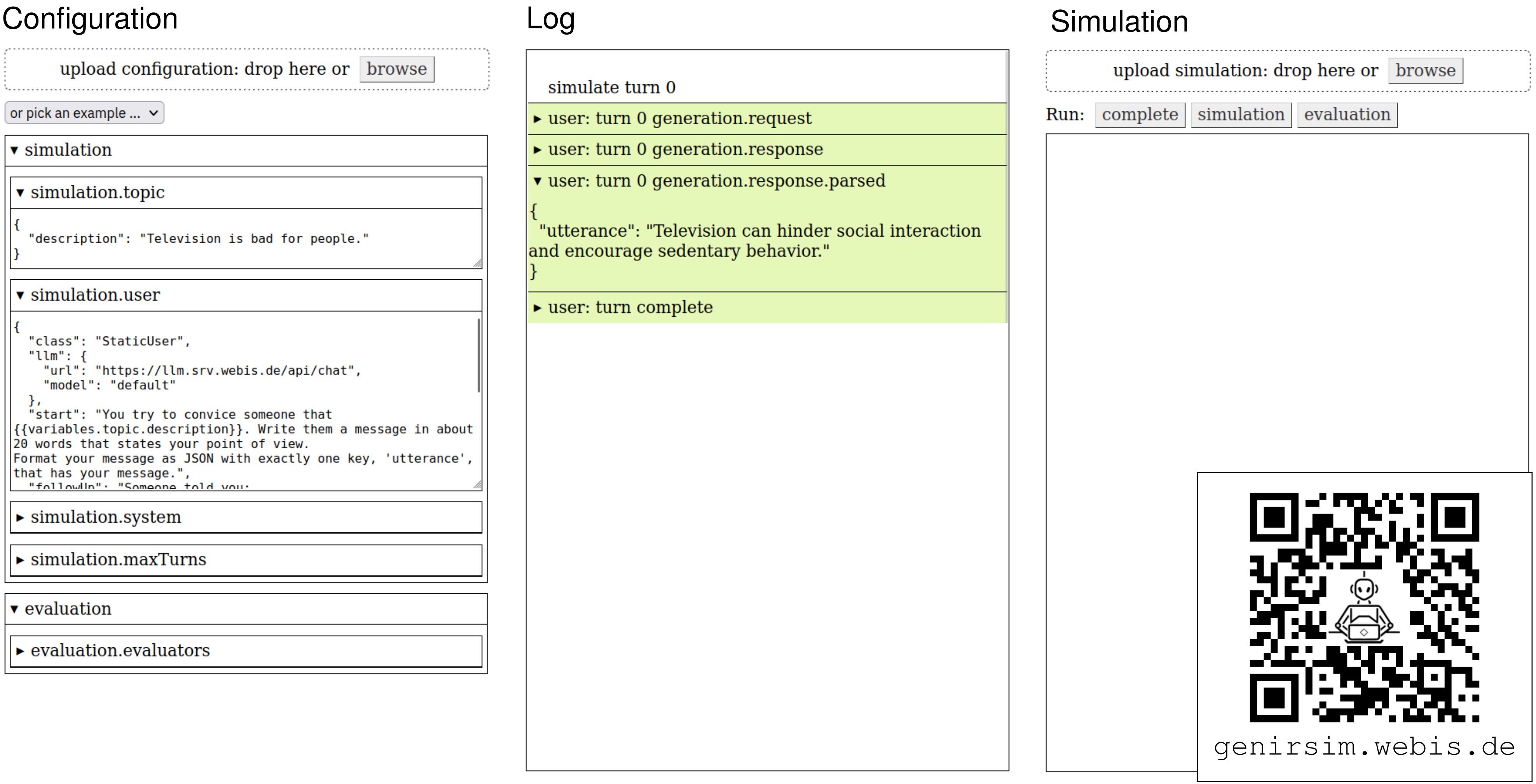
```
{
  "class": "StaticUser",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "start": "You try to convice someone that
{{variables.topic.description}}. Write them a message in about
20 words that states your point of view.
Format your message as JSON with exactly one key, 'utterance',
that has your message.",
  "followUp": "Someone told you:
```

```
[
  {
    "role": "user",
    "content": "You try to convice someone that
Television is bad for people.. Write them a message in
about 20 words that states your point of view.

Format your message as JSON with exactly one key,
'utterance', that has your message.

"
  }
]
```

Simulation
upload simulation: drop here or browse
Run: complete simulation evaluation

genirsim.webis.de

GenIRSim



Configuration

```
{
  "description": "Television is bad for people."
}
```

```
{
  "class": "StaticUser",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "start": "You try to convice someone that
{{variables.topic.description}}. Write them a message in about
20 words that states your point of view.
Format your message as JSON with exactly one key, 'utterance',
that has your message.",
  "followUp": "Someone told you:
```

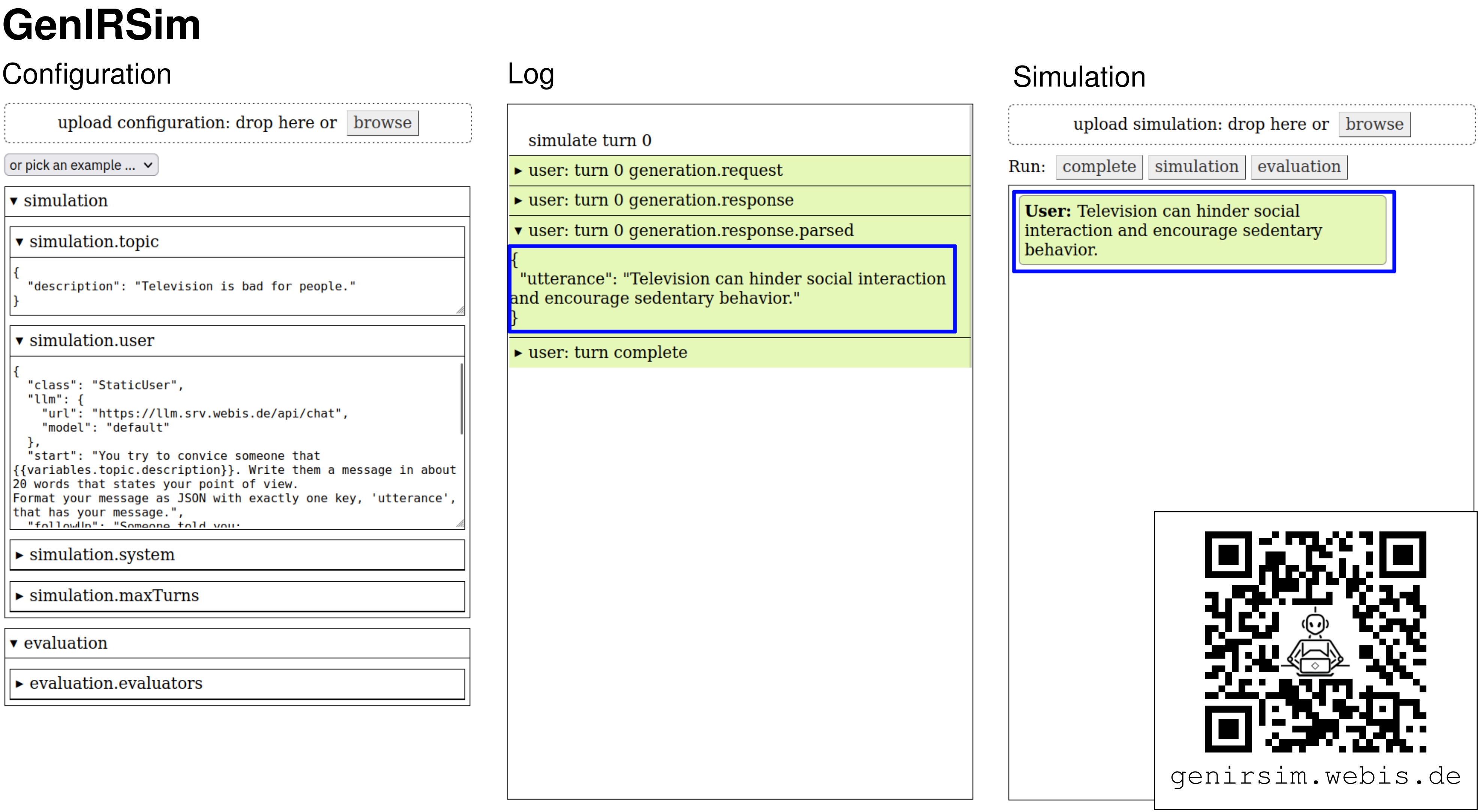
Log

```
simulate turn 0
▸ user: turn 0 generation.request
▸ user: turn 0 generation.response
▾ user: turn 0 generation.response.parsed
{
  "utterance": "Television can hinder social interaction and encourage sedentary behavior."
}
▸ user: turn complete
```

Simulation

upload simulation: drop here or browse
Run: complete simulation evaluation

genirsim.webis.de



GenIRSim

Configuration

```
{
  "description": "Television is bad for people."
}
```

```
{
  "class": "StaticUser",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "start": "You try to convice someone that
{{variables.topic.description}}. Write them a message in about
20 words that states your point of view.
Format your message as JSON with exactly one key, 'utterance',
that has your message.",
  "followUp": "Someone told you:
```

Log

```
simulate turn 0
▸ user: turn 0 generation.request
▸ user: turn 0 generation.response
▾ user: turn 0 generation.response.parsed
{
  "utterance": "Television can hinder social interaction
and encourage sedentary behavior."
}
▸ user: turn complete
```

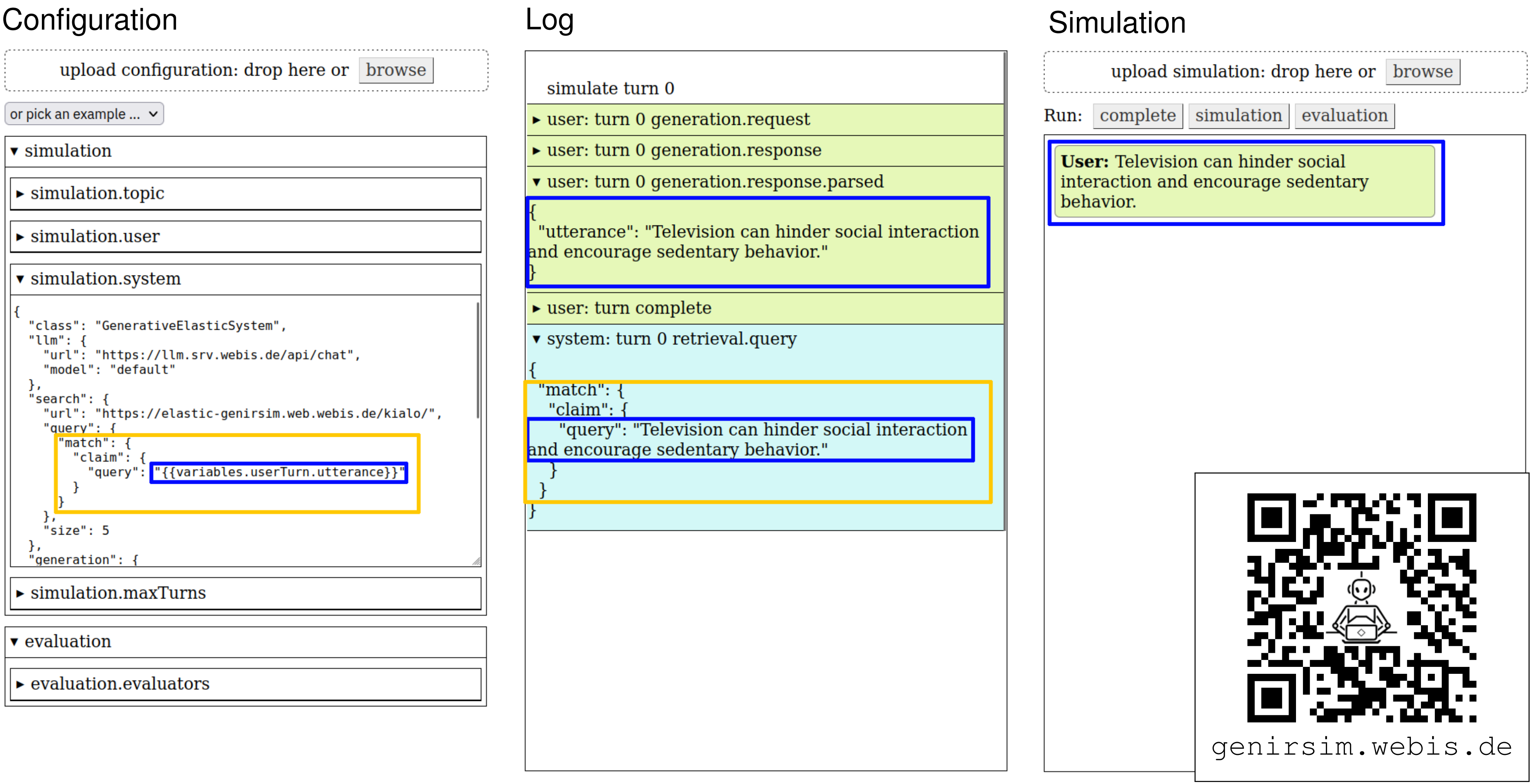
Simulation

upload simulation: drop here or browse
Run: complete simulation evaluation

User: Television can hinder social interaction and encourage sedentary behavior.

genirsim.webis.de

GenIRSim



Configuration

▾ simulation

▸ simulation.topic

▸ simulation.user

▾ simulation.system

```
{
  "class": "GenerativeElasticSystem",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "search": {
    "url": "https://elastic-genirsim.web.webis.de/kialo/",
    "query": {
      "match": {
        "claim": {
          "query": "{{variables.userTurn.utterance}}"
        }
      }
    },
    "size": 5
  },
  "generation": {
```

▸ simulation.maxTurns

▾ evaluation

▸ evaluation.evaluators

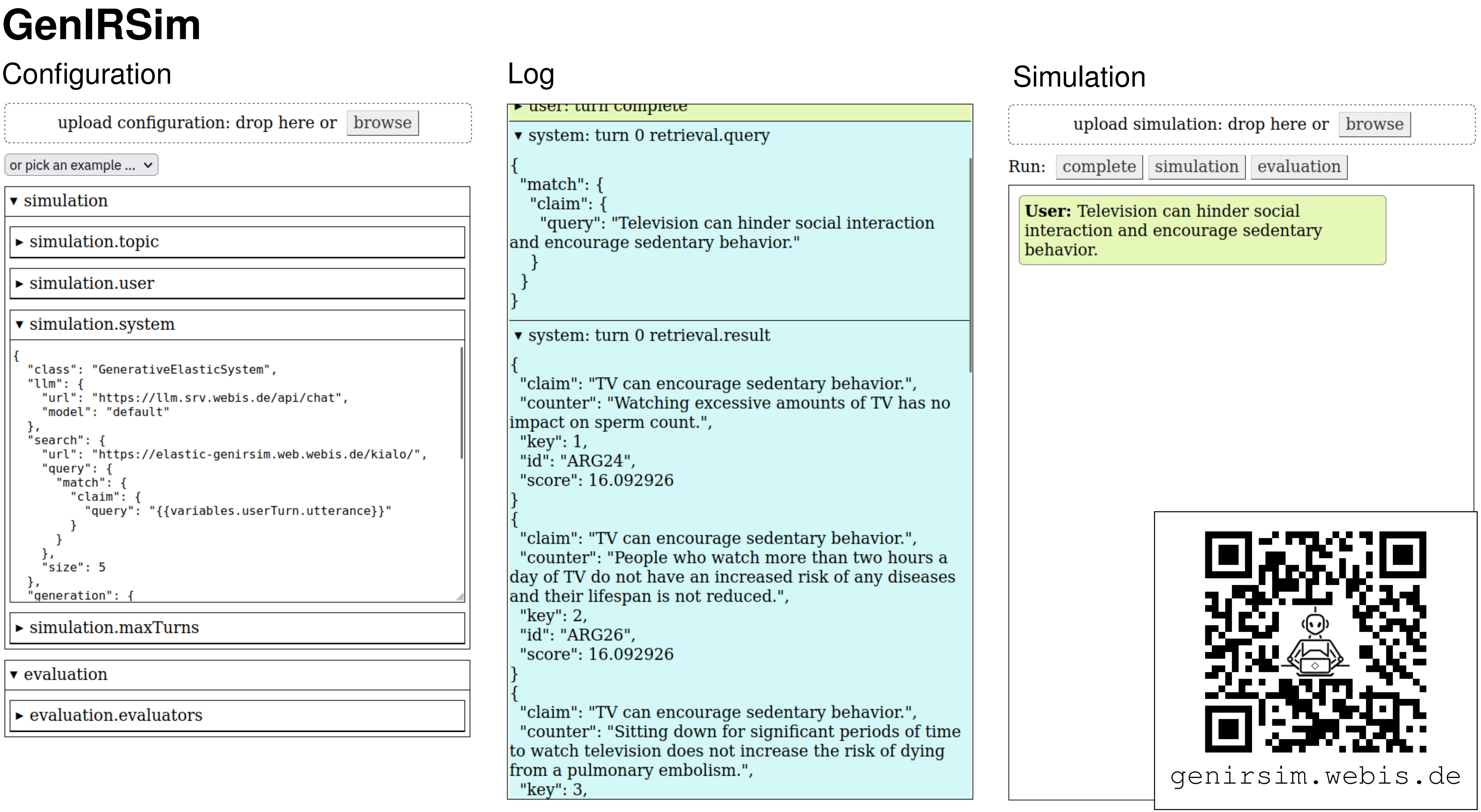
Log

```
simulate turn 0
▸ user: turn 0 generation.request
▸ user: turn 0 generation.response
▾ user: turn 0 generation.response.parsed
{
  "utterance": "Television can hinder social interaction and encourage sedentary behavior."
}
▸ user: turn complete
▾ system: turn 0 retrieval.query
{
  "match": {
    "claim": {
      "query": "Television can hinder social interaction and encourage sedentary behavior."
    }
  }
}
```

Simulation

User: Television can hinder social interaction and encourage sedentary behavior.

genirsim.webis.de



GenIRSim

Configuration

upload configuration: drop here or browse

or pick an example ...

▾ simulation

▸ simulation.topic

▸ simulation.user

▾ simulation.system

```
{
  "class": "GenerativeElasticSystem",
  "llm": {
    "url": "https://llm.srv.webis.de/api/chat",
    "model": "default"
  },
  "search": {
    "url": "https://elastic-genirsim.web.webis.de/kialo/",
    "query": {
      "match": {
        "claim": {
          "query": "{{variables.userTurn.utterance}}"
        }
      }
    },
    "size": 5
  },
  "generation": {
```

▸ simulation.maxTurns

▾ evaluation

▸ evaluation.evaluators

Log

▸ user: turn complete

▾ system: turn 0 retrieval.query

```
{
  "match": {
   "claim": {
     "query": "Television can hinder social interaction and encourage sedentary behavior."
   }
  }
}
```

▾ system: turn 0 retrieval.result

```
{
  "claim": "TV can encourage sedentary behavior.",
  "counter": "Watching excessive amounts of TV has no impact on sperm count.",
  "key": 1,
  "id": "ARG24",
  "score": 16.092926
}
{
  "claim": "TV can encourage sedentary behavior.",
  "counter": "People who watch more than two hours a day of TV do not have an increased risk of any diseases and their lifespan is not reduced.",
  "key": 2,
  "id": "ARG26",
  "score": 16.092926
}
{
  "claim": "TV can encourage sedentary behavior.",
  "counter": "Sitting down for significant periods of time to watch television does not increase the risk of dying from a pulmonary embolism.",
  "key": 3,
```

Simulation

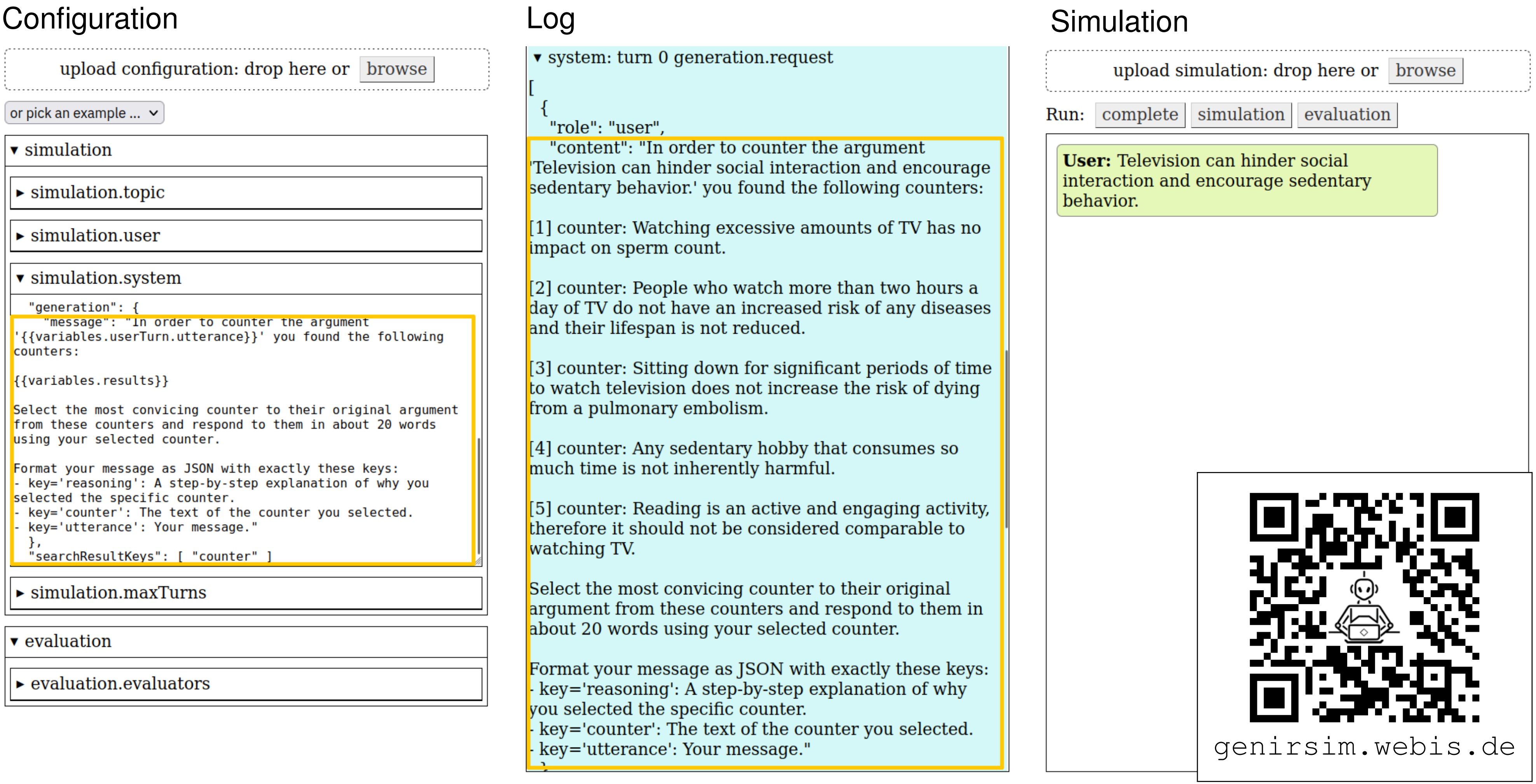
upload simulation: drop here or browse

Run: complete simulation evaluation

User: Television can hinder social interaction and encourage sedentary behavior.

genirsim.webis.de

GenIRSim



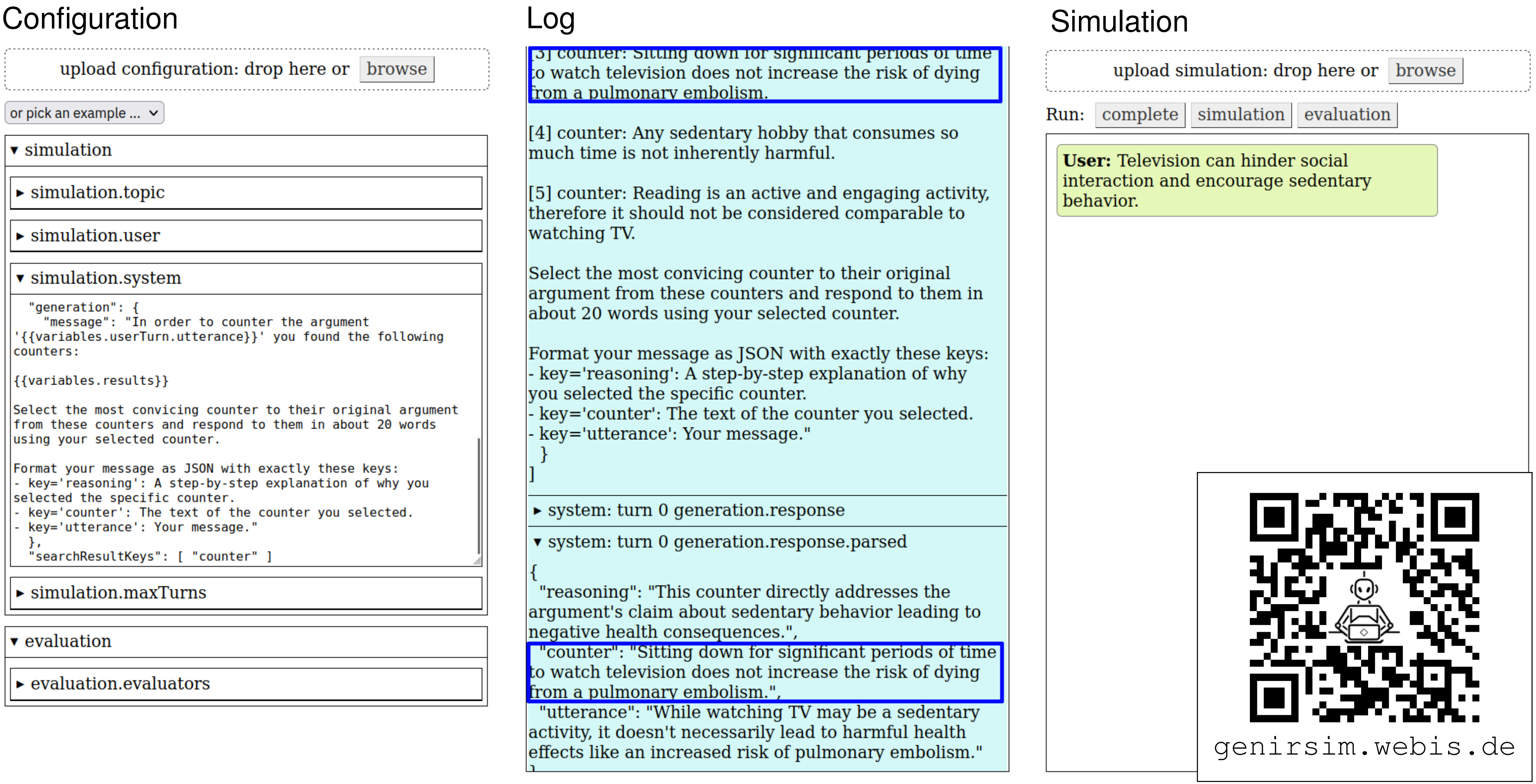
```
  "generation": {
    "message": "In order to counter the argument
'{{variables.userTurn.utterance}}' you found the following
counters:

{{variables.results}}

Select the most convicing counter to their original argument
from these counters and respond to them in about 20 words
using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why you
selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
  },
  "searchResultKeys": [ "counter" ]
```

```
[
  {
    "role": "user",
    "content": "In order to counter the argument
'Television can hinder social interaction and encourage
sedentary behavior.' you found the following counters:

[1] counter: Watching excessive amounts of TV has no
impact on sperm count.

[2] counter: People who watch more than two hours a
day of TV do not have an increased risk of any diseases
and their lifespan is not reduced.

[3] counter: Sitting down for significant periods of time
to watch television does not increase the risk of dying
from a pulmonary embolism.

[4] counter: Any sedentary hobby that consumes so
much time is not inherently harmful.

[5] counter: Reading is an active and engaging activity,
therefore it should not be considered comparable to
watching TV.

Select the most convicing counter to their original
argument from these counters and respond to them in
about 20 words using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why
you selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
```

User: Television can hinder social interaction and encourage sedentary behavior.

genirsim.webis.de

GenIRSim


Configuration
upload configuration: drop here or browse
or pick an example ...
Log
Simulation
upload simulation: drop here or browse
Run: complete simulation evaluation

▾ simulation

▸ simulation.topic

▸ simulation.user

▾ simulation.system

```
  "generation": {
    "message": "In order to counter the argument
'{{variables.userTurn.utterance}}' you found the following
counters:

{{variables.results}}

Select the most convicing counter to their original argument
from these counters and respond to them in about 20 words
using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why you
selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
  },
  "searchResultKeys": [ "counter" ]
```

▸ simulation.maxTurns

▾ evaluation

▸ evaluation.evaluators

```
[3] counter: Sitting down for significant periods of time
to watch television does not increase the risk of dying
from a pulmonary embolism.

[4] counter: Any sedentary hobby that consumes so
much time is not inherently harmful.

[5] counter: Reading is an active and engaging activity,
therefore it should not be considered comparable to
watching TV.

Select the most convicing counter to their original
argument from these counters and respond to them in
about 20 words using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why
you selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
  }
]
```

▸ system: turn 0 generation.response

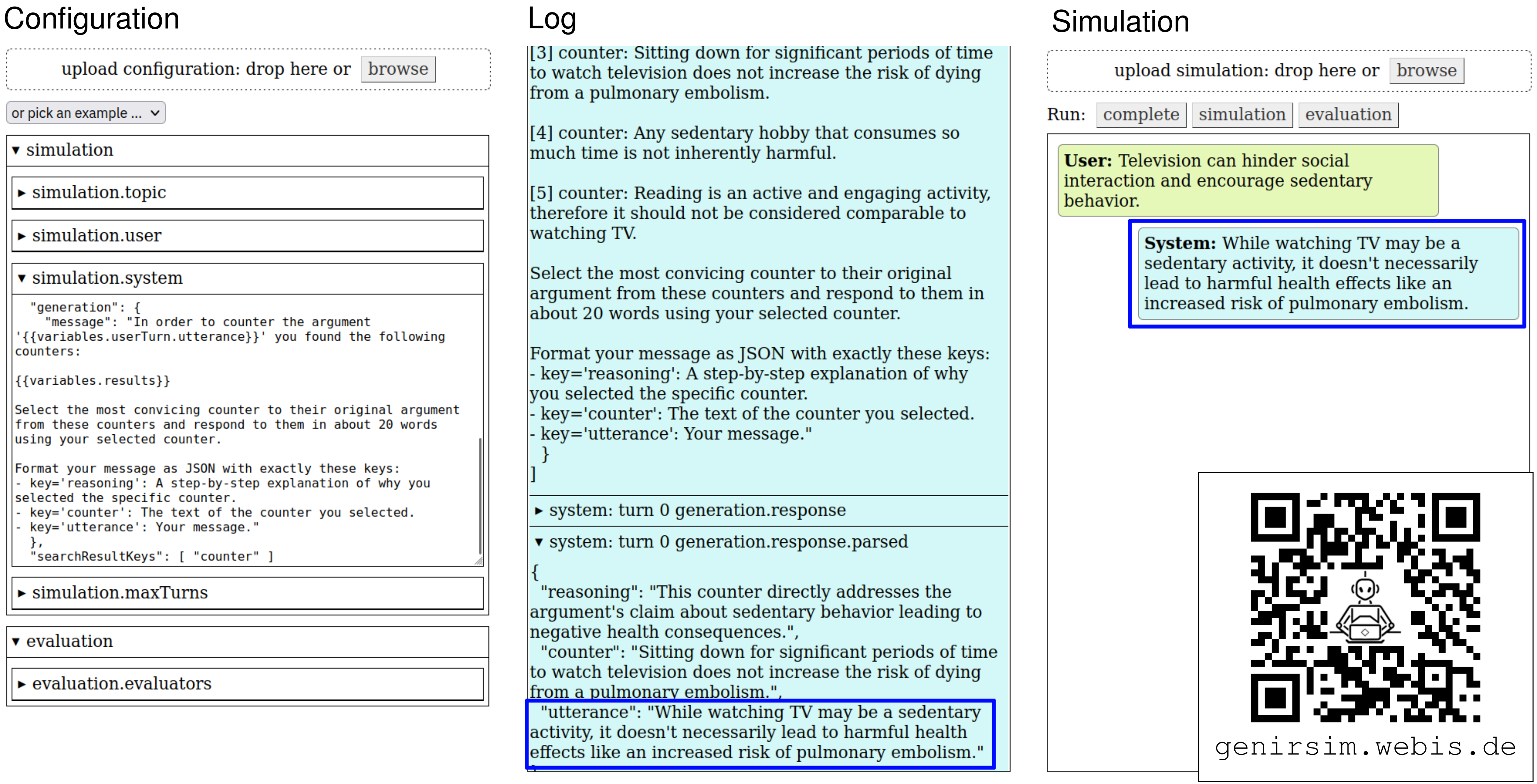
▾ system: turn 0 generation.response.parsed

```
{
  "reasoning": "This counter directly addresses the
argument's claim about sedentary behavior leading to
negative health consequences.",
  "counter": "Sitting down for significant periods of time
to watch television does not increase the risk of dying
from a pulmonary embolism.",
  "utterance": "While watching TV may be a sedentary
activity, it doesn't necessarily lead to harmful health
effects like an increased risk of pulmonary embolism."
```

User: Television can hinder social interaction and encourage sedentary behavior.

genirsim.webis.de

GenIRSim



```
  "generation": {
    "message": "In order to counter the argument
'{{variables.userTurn.utterance}}' you found the following
counters:

{{variables.results}}

Select the most convicing counter to their original argument
from these counters and respond to them in about 20 words
using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why you
selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
  },
  "searchResultKeys": [ "counter" ]
```

```
[3] counter: Sitting down for significant periods of time
to watch television does not increase the risk of dying
from a pulmonary embolism.

[4] counter: Any sedentary hobby that consumes so
much time is not inherently harmful.

[5] counter: Reading is an active and engaging activity,
therefore it should not be considered comparable to
watching TV.

Select the most convicing counter to their original
argument from these counters and respond to them in
about 20 words using your selected counter.

Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why
you selected the specific counter.
- key='counter': The text of the counter you selected.
- key='utterance': Your message."
  }
]
```

▸ system: turn 0 generation.response

▾ system: turn 0 generation.response.parsed

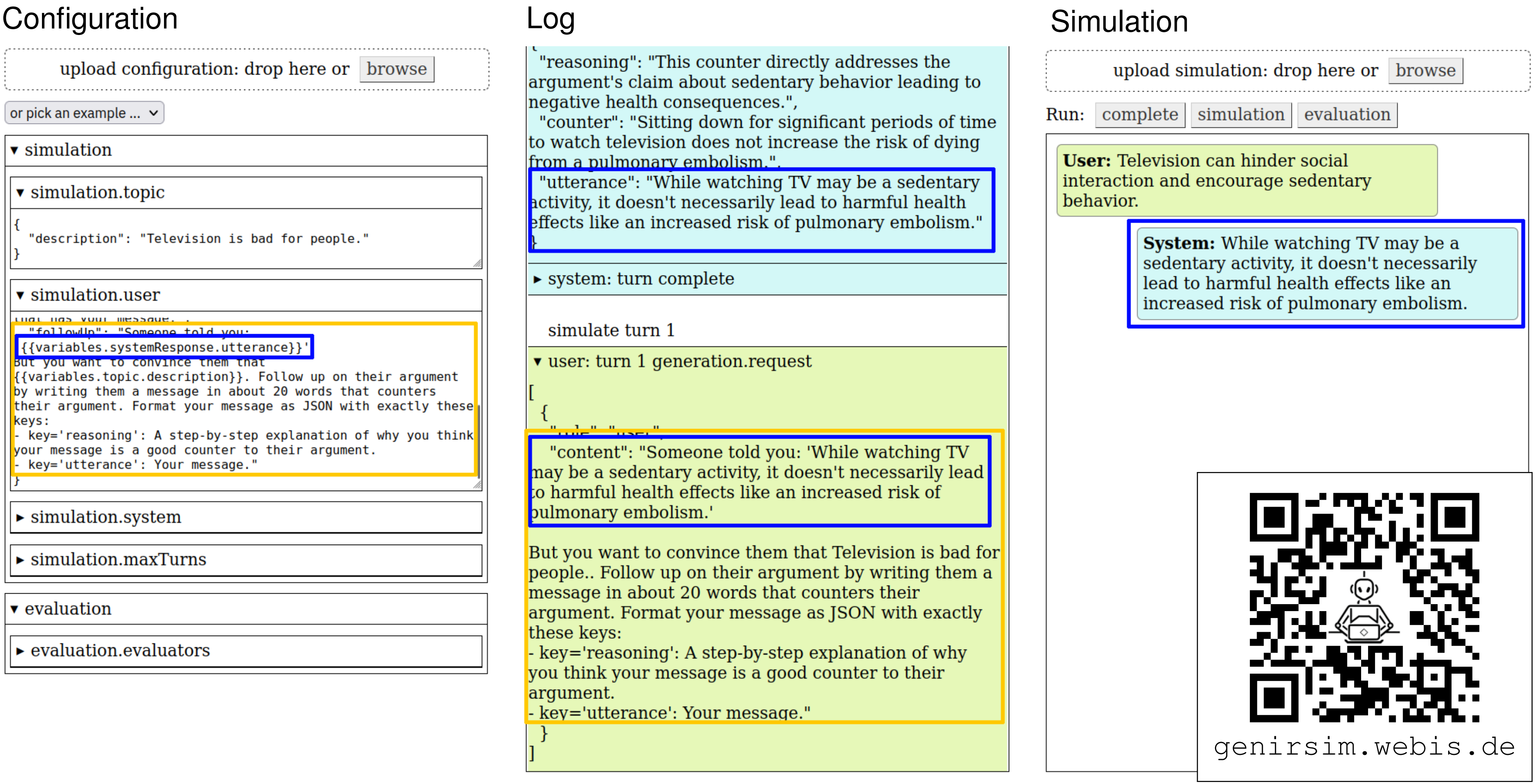
```
{
  "reasoning": "This counter directly addresses the
argument's claim about sedentary behavior leading to
negative health consequences.",
  "counter": "Sitting down for significant periods of time
to watch television does not increase the risk of dying
from a pulmonary embolism.",
  "utterance": "While watching TV may be a sedentary
activity, it doesn't necessarily lead to harmful health
effects like an increased risk of pulmonary embolism."
```

User: Television can hinder social interaction and encourage sedentary behavior.

System: While watching TV may be a sedentary activity, it doesn't necessarily lead to harmful health effects like an increased risk of pulmonary embolism.

genirsim.webis.de

GenIRSim



Configuration

```
{
  "description": "Television is bad for people."
}
```

```
  "followUp": "Someone told you:
{{variables.systemResponse.utterance}}'
But you want to convince them that
{{variables.topic.description}}. Follow up on their argument
by writing them a message in about 20 words that counters
their argument. Format your message as JSON with exactly these
keys:
- key='reasoning': A step-by-step explanation of why you think
your message is a good counter to their argument.
- key='utterance': Your message."
}
```

Log

```
  "reasoning": "This counter directly addresses the argument's claim about sedentary behavior leading to negative health consequences.",
  "counter": "Sitting down for significant periods of time to watch television does not increase the risk of dying from a pulmonary embolism.",
  "utterance": "While watching TV may be a sedentary activity, it doesn't necessarily lead to harmful health effects like an increased risk of pulmonary embolism."
}
```

simulate turn 1

```
[
  {
    "role": "user",
    "content": "Someone told you: 'While watching TV may be a sedentary activity, it doesn't necessarily lead to harmful health effects like an increased risk of pulmonary embolism.'

But you want to convince them that Television is bad for people.. Follow up on their argument by writing them a message in about 20 words that counters their argument. Format your message as JSON with exactly these keys:
- key='reasoning': A step-by-step explanation of why you think your message is a good counter to their argument.
- key='utterance': Your message."
  }
]
```

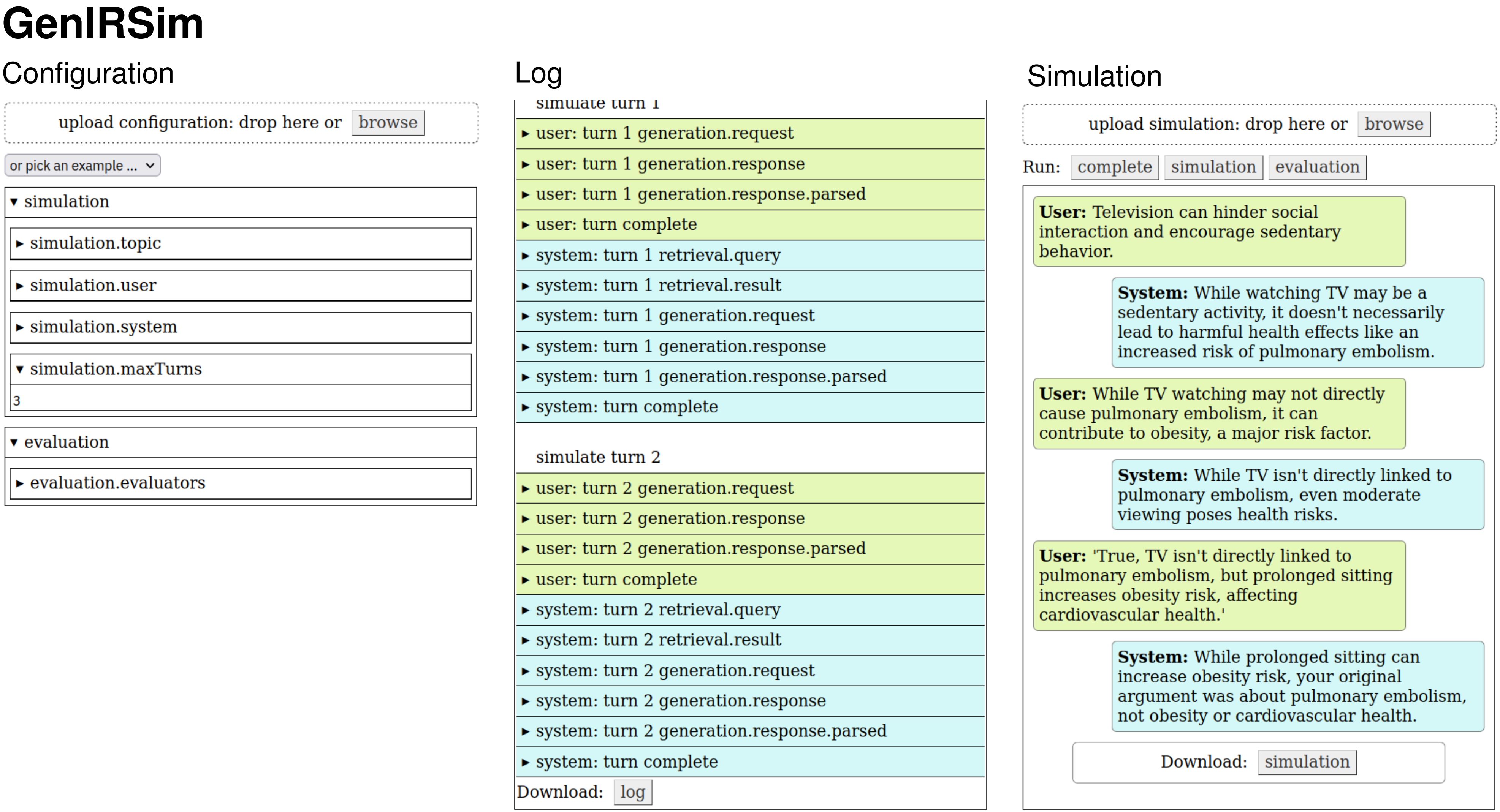
Simulation

upload simulation: drop here or browse
Run: complete simulation evaluation

User: Television can hinder social interaction and encourage sedentary behavior.

System: While watching TV may be a sedentary activity, it doesn't necessarily lead to harmful health effects like an increased risk of pulmonary embolism.

genirsim.webis.de



- ▾ simulation
  - ▸ simulation.topic
  - ▸ simulation.user
  - ▸ simulation.system
  - ▾ simulation.maxTurns
    3
- ▾ evaluation
  - ▸ evaluation.evaluators

```
simulate turn 1
▸ user: turn 1 generation.request
▸ user: turn 1 generation.response
▸ user: turn 1 generation.response.parsed
▸ user: turn complete
▸ system: turn 1 retrieval.query
▸ system: turn 1 retrieval.result
▸ system: turn 1 generation.request
▸ system: turn 1 generation.response
▸ system: turn 1 generation.response.parsed
▸ system: turn complete

simulate turn 2
▸ user: turn 2 generation.request
▸ user: turn 2 generation.response
▸ user: turn 2 generation.response.parsed
▸ user: turn complete
▸ system: turn 2 retrieval.query
▸ system: turn 2 retrieval.result
▸ system: turn 2 generation.request
▸ system: turn 2 generation.response
▸ system: turn 2 generation.response.parsed
▸ system: turn complete
```

User: Television can hinder social interaction and encourage sedentary behavior.

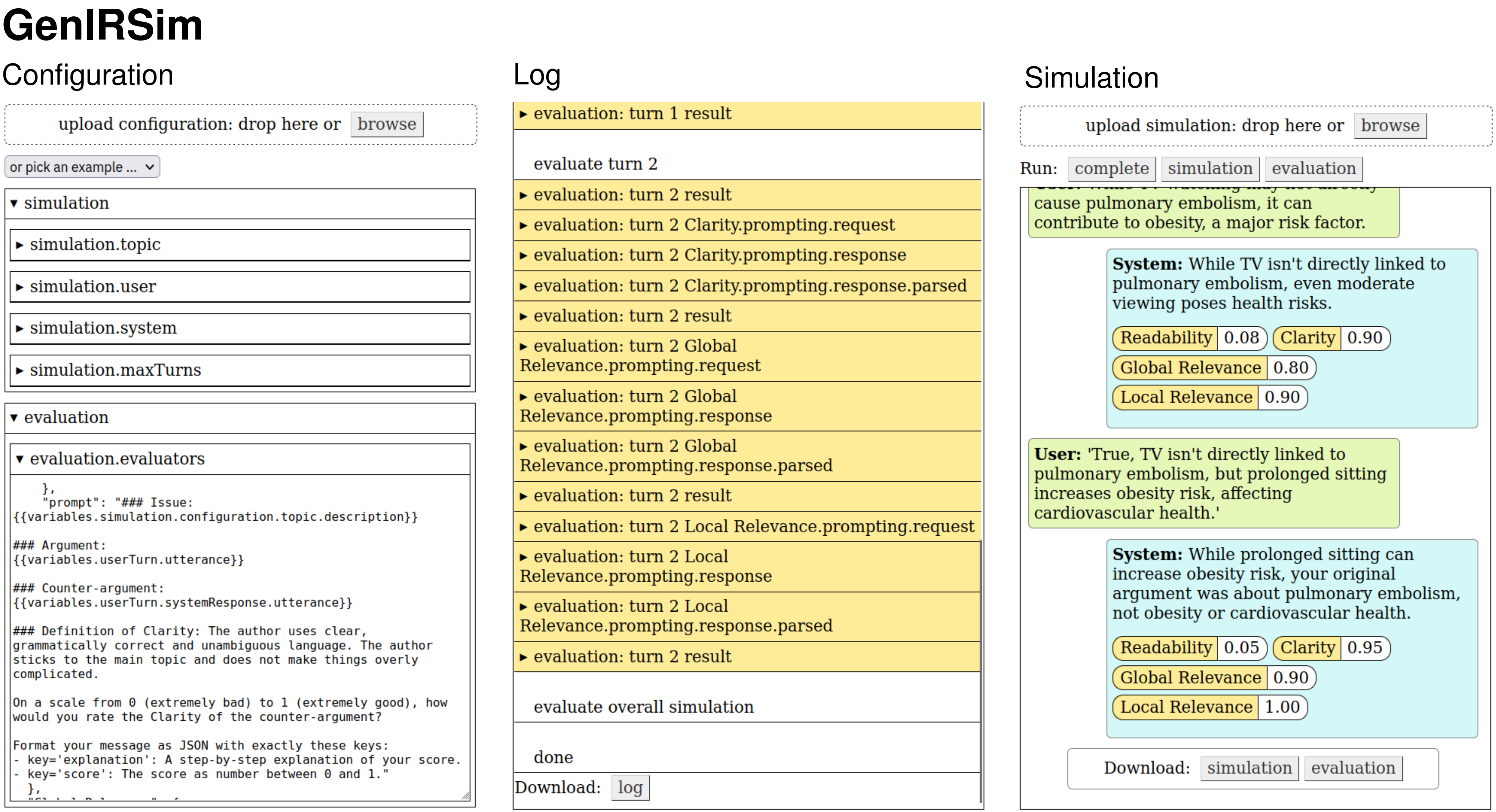
System: While watching TV may be a sedentary activity, it doesn't necessarily lead to harmful health effects like an increased risk of pulmonary embolism.

User: While TV watching may not directly cause pulmonary embolism, it can contribute to obesity, a major risk factor.

System: While TV isn't directly linked to pulmonary embolism, even moderate viewing poses health risks.

User: 'True, TV isn't directly linked to pulmonary embolism, but prolonged sitting increases obesity risk, affecting cardiovascular health.'

System: While prolonged sitting can increase obesity risk, your original argument was about pulmonary embolism, not obesity or cardiovascular health.



GenIRSim

Configuration

upload configuration: drop here or

▾ simulation

▸ simulation.topic

▸ simulation.user

▸ simulation.system

▸ simulation.maxTurns

▾ evaluation

▾ evaluation.evaluators

```
    },
    "prompt": "### Issue:
{{variables.simulation.configuration.topic.description}}

### Argument:
{{variables.userTurn.utterance}}

### Counter-argument:
{{variables.userTurn.systemResponse.utterance}}

### Definition of Clarity: The author uses clear,
grammatically correct and unambiguous language. The author
sticks to the main topic and does not make things overly
complicated.

On a scale from 0 (extremely bad) to 1 (extremely good), how
would you rate the Clarity of the counter-argument?

Format your message as JSON with exactly these keys:
- key='explanation': A step-by-step explanation of your score.
- key='score': The score as number between 0 and 1."
  },
```

Log

▸ evaluation: turn 1 result

evaluate turn 2

▸ evaluation: turn 2 result

▸ evaluation: turn 2 Clarity.prompting.request

▸ evaluation: turn 2 Clarity.prompting.response

▸ evaluation: turn 2 Clarity.prompting.response.parsed

▸ evaluation: turn 2 result

▸ evaluation: turn 2 Global Relevance.prompting.request

▸ evaluation: turn 2 Global Relevance.prompting.response

▸ evaluation: turn 2 Global Relevance.prompting.response.parsed

▸ evaluation: turn 2 result

▸ evaluation: turn 2 Local Relevance.prompting.request

▸ evaluation: turn 2 Local Relevance.prompting.response

▸ evaluation: turn 2 Local Relevance.prompting.response.parsed

▸ evaluation: turn 2 result

evaluate overall simulation

done

Download:

Simulation

upload simulation: drop here or

Run:

cause pulmonary embolism, it can contribute to obesity, a major risk factor.

System: While TV isn't directly linked to pulmonary embolism, even moderate viewing poses health risks.

Readability 0.08 | Clarity 0.90
Global Relevance 0.80
Local Relevance 0.90

User: 'True, TV isn't directly linked to pulmonary embolism, but prolonged sitting increases obesity risk, affecting cardiovascular health.'

System: While prolonged sitting can increase obesity risk, your original argument was about pulmonary embolism, not obesity or cardiovascular health.

Readability 0.05 | Clarity 0.95
Global Relevance 0.90
Local Relevance 1.00

Download:

Exploring the Gen-IR Simulation and Evaluation Space

Six Directions for Future Research

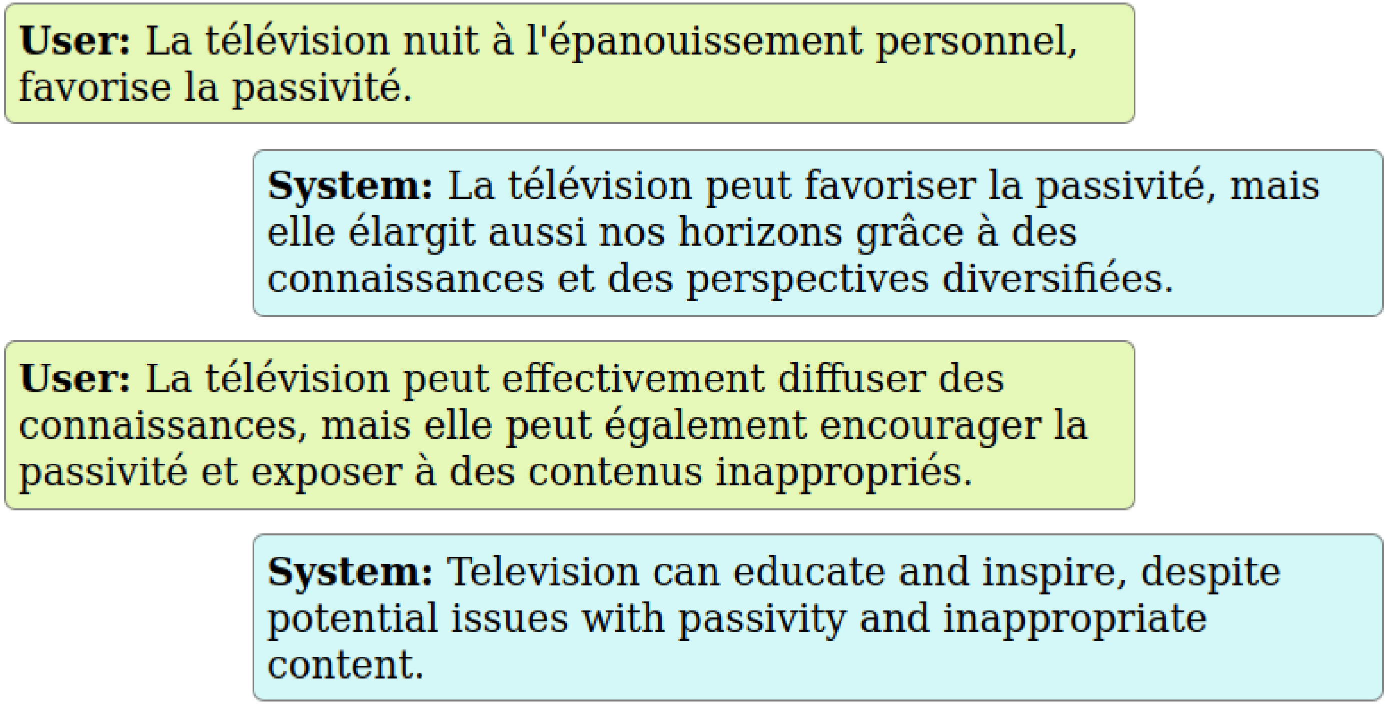
- ❑ User information and knowledge
  How to model users?
- ❑ User selection
  Whom to simulate?
- ❑ Multilingualism
  Is user simulation restricted to English?
- ❑ User model updates
  How to model that users acquire new knowledge during search?
- ❑ Evaluation aspects
  What makes a “good” interaction with a Gen-IR system?
- ❑ “Thought” processes
  Can we ask the simulated user what it “thinks?”

Exploring the Gen-IR Simulation and Evaluation Space

Is User Simulation Restricted to English?

Addition to user prompts: *Answer in French*

Addition to system prompts: *Answer in the same language as the argument*



User: La télévision nuit à l'épanouissement personnel, favorise la passivité.

System: La télévision peut favoriser la passivité, mais elle élargit aussi nos horizons grâce à des connaissances et des perspectives diversifiées.

User: La télévision peut effectivement diffuser des connaissances, mais elle peut également encourager la passivité et exposer à des contenus inappropriés.

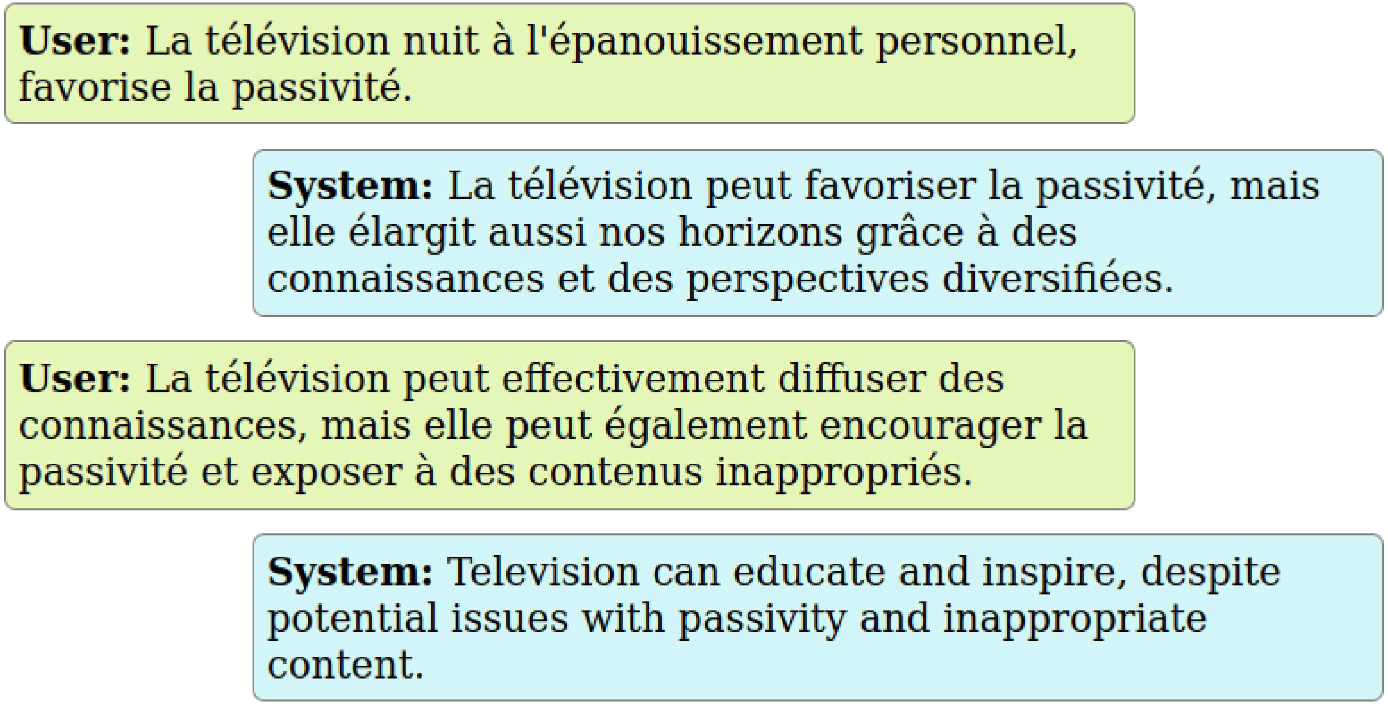
System: Television can educate and inspire, despite potential issues with passivity and inappropriate content.

Exploring the Gen-IR Simulation and Evaluation Space

Is User Simulation Restricted to English?

Addition to user prompts: *Answer in French*

Addition to system prompts: *Answer in the same language as the argument*



User: La télévision nuit à l'épanouissement personnel, favorise la passivité.

System: La télévision peut favoriser la passivité, mais elle élargit aussi nos horizons grâce à des connaissances et des perspectives diversifiées.

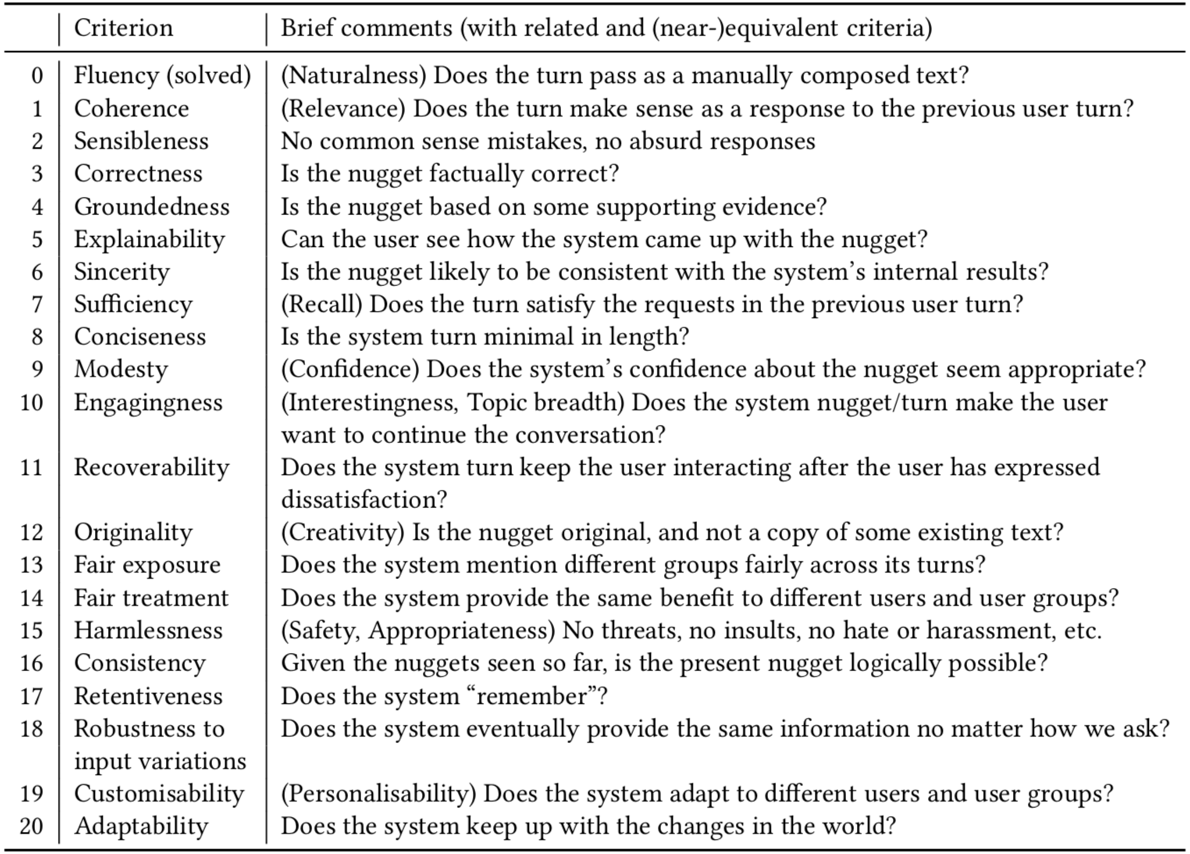
User: La télévision peut effectivement diffuser des connaissances, mais elle peut également encourager la passivité et exposer à des contenus inappropriés.

System: Television can educate and inspire, despite potential issues with passivity and inappropriate content.

(Retrieval actually failed; LLM answered anyway)

Exploring the Gen-IR Simulation and Evaluation Space

Evaluation Aspects



	Criterion	Brief comments (with related and (near-)equivalent criteria)
0	Fluency (solved)	(Naturalness) Does the turn pass as a manually composed text?
1	Coherence	(Relevance) Does the turn make sense as a response to the previous user turn?
2	Sensibleness	No common sense mistakes, no absurd responses
3	Correctness	Is the nugget factually correct?
4	Groundedness	Is the nugget based on some supporting evidence?
5	Explainability	Can the user see how the system came up with the nugget?
6	Sincerity	Is the nugget likely to be consistent with the system's internal results?
7	Sufficiency	(Recall) Does the turn satisfy the requests in the previous user turn?
8	Conciseness	Is the system turn minimal in length?
9	Modesty	(Confidence) Does the system's confidence about the nugget seem appropriate?
10	Engagingness	(Interestingness, Topic breadth) Does the system nugget/turn make the user want to continue the conversation?
11	Recoverability	Does the system turn keep the user interacting after the user has expressed dissatisfaction?
12	Originality	(Creativity) Is the nugget original, and not a copy of some existing text?
13	Fair exposure	Does the system mention different groups fairly across its turns?
14	Fair treatment	Does the system provide the same benefit to different users and user groups?
15	Harmlessness	(Safety, Appropriateness) No threats, no insults, no hate or harassment, etc.
16	Consistency	Given the nuggets seen so far, is the present nugget logically possible?
17	Retentiveness	Does the system "remember"?
18	Robustness to input variations	Does the system eventually provide the same information no matter how we ask?
19	Customisability	(Personalisability) Does the system adapt to different users and user groups?
20	Adaptability	Does the system keep up with the changes in the world?

SWAN: A Generic Framework for Auditing Textual Conversational Systems. Tetsuya Sakai. 2023

Exploring the Gen-IR Simulation and Evaluation Space

Can We Ask the Simulated User What it “Thinks?”

Formatting instructions in user prompt (chain-of-thought-like):

```
[...]  write a message [...]  to request the information you are lacking.
Format your message as JSON with exactly these keys:
- key='expectation':  A description of what you expect the system's answer to your
message to contain.
- key='utterance':  Your short message.
```

Exploring the Gen-IR Simulation and Evaluation Space

Can We Ask the Simulated User What it “Thinks?”

Formatting instructions in user prompt (chain-of-thought-like):

```
[...] write a message [...] to request the information you are lacking.
Format your message as JSON with exactly these keys:
- key=’expectation’: A description of what you expect the system’s answer to your
message to contain.
- key=’utterance’: Your short message.
```

Evaluation prompt:

```
### Expectation:
{{variables.userTurn.expectation}}

### Response:
{{variables.userTurn.systemResponse.utterance}}

### Instruction:
Score from 0 (extremely bad) to 1 (extremely good) whether the response matches the
expectation.
```

Interested? Be RAD @ Touché’25!

Shared Task: Retrieval-Augmented Debating

Subtask 1: Retrieve and respond with counterarguments and evidence in simulated debates
Subtask 2: Automate the evaluation of such systems

User – U_1: Claim statement

S_1: Supposed to attack U_1 – System

User – U_2: Attacks S_1

S_2: Supposed to respond to U_2 – System

User – U_3: Attacks S_1 or S_2

S_3: Supposed to respond to U_3 – System

Simulated debate procedure for evaluation

touche.webis.de

genirsim.webis.de