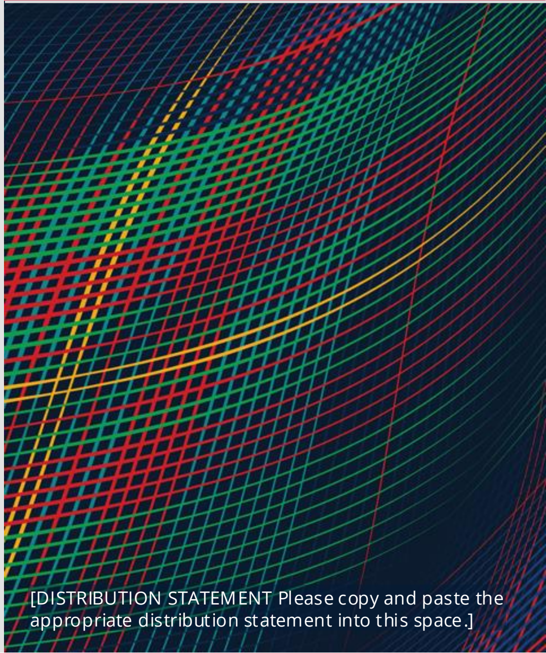
Carnegie Mellon University
Software Engineering Institute

Discourse analysis as a diagnostic lens:

Untangling some of the riddles complicating LLM evaluations

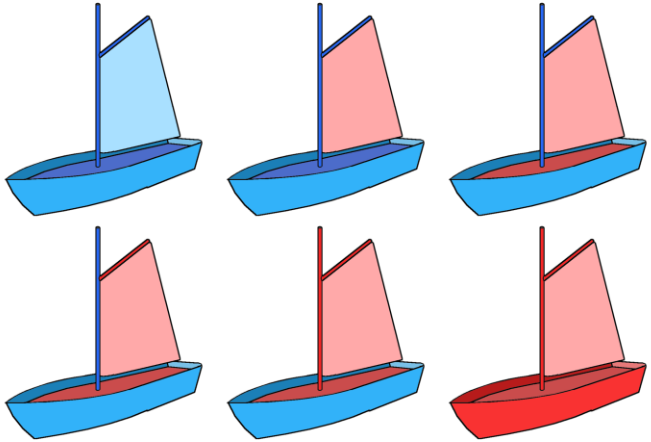
SEPTEMBER, 2025

Samantha Finkelstein, PhD
Sr. Human-Centered AI Research Scientist,



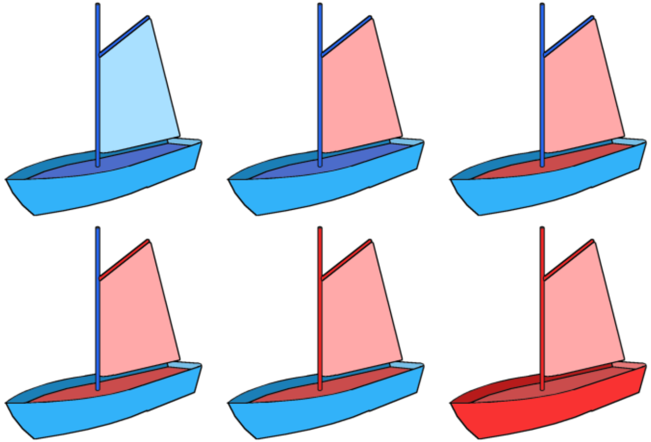
[DISTRIBUTION STATEMENT Please copy and paste the appropriate distribution statement into this space.]

What are we trying to evaluate?



What are we trying to evaluate?

The Ship of Theseus is a philosophy paradox:

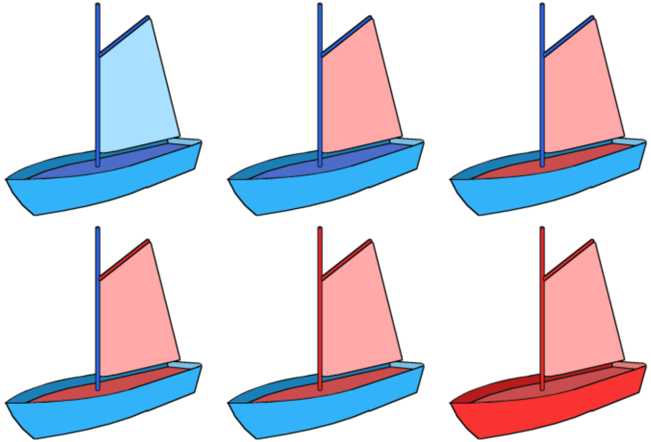


If we replace one plank in the ship of Theseus, is it the same ship?

What if we've replaced every plank?

What are we trying to evaluate?

The Ship of Theseus is a philosophy paradox:



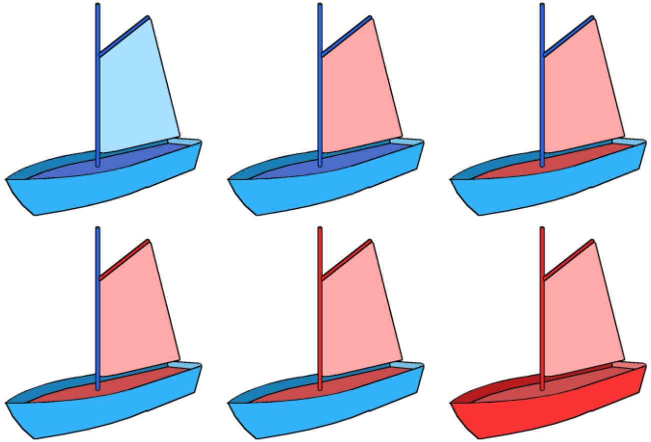
If we replace one plank in the ship of Theseus, is it the same ship?

What if we've replaced every plank?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

What are we trying to evaluate?

The Ship of Theseus is a philosophy paradox:



If we replace one plank in the ship of Theseus, is it the same ship?

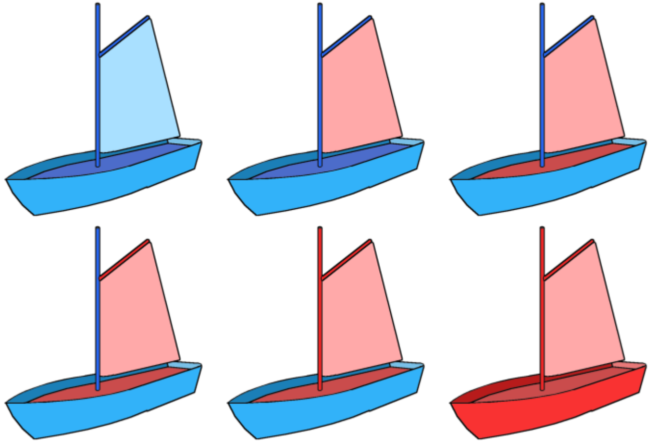
What if we've replaced every plank?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

The Ship of Theseus is a ~~philosophy paradox~~ definition trick:



If we replace one plank in the ship of Theseus, is it the same ship?

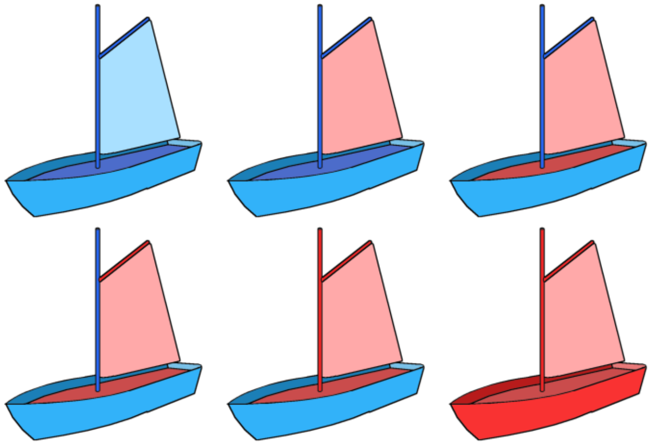
What if we've replaced every plank?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



If we replace one plank in the ship of Theseus, is it the same ship?

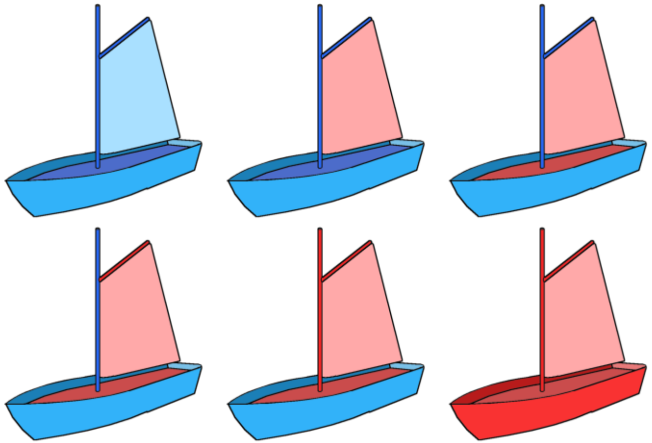
What if we've replaced every plank?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



Are LLMs trustworthy?

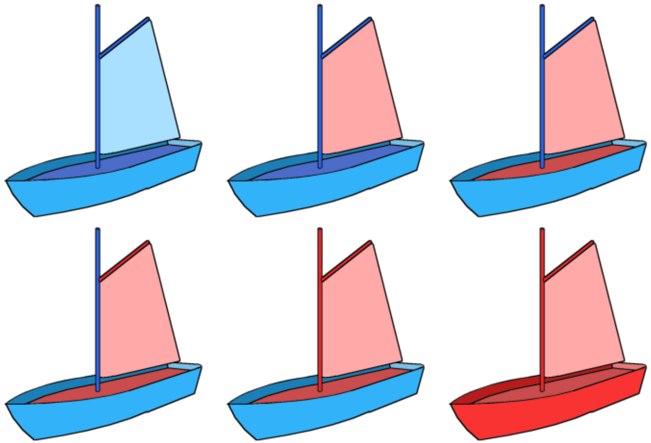
Are they reliable? Are they safe?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



Are LLMs trustworthy?

What do we mean by trustworthy?

Are they reliable? Are they safe?

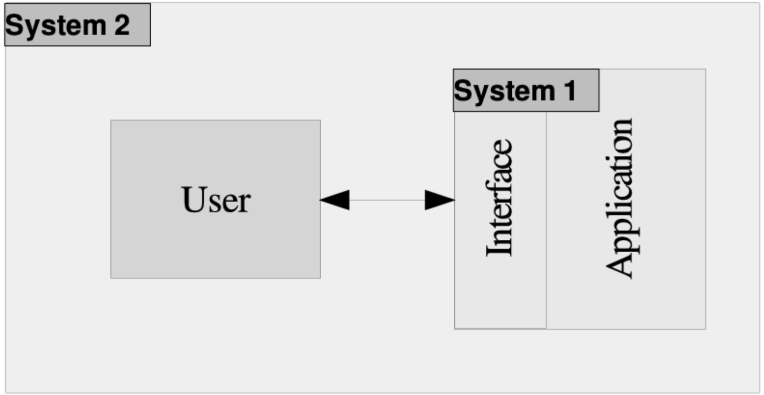
What do we mean by reliable? By safe?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



Are LLMs trustworthy?

What do we mean by trustworthy?

Are they reliable? Are they safe?

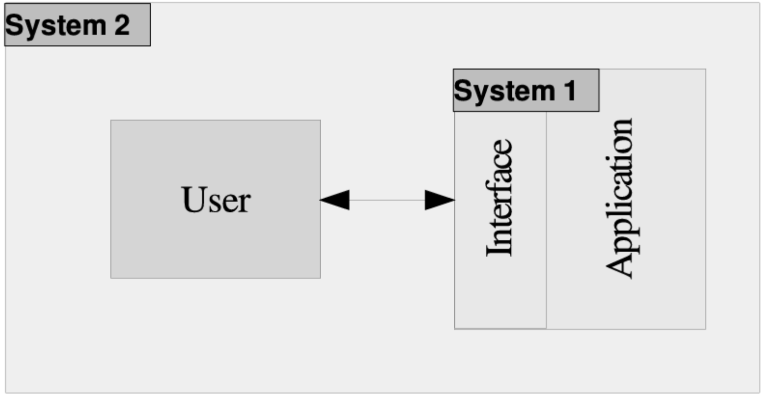
What do we mean by reliable? By safe?

- If "**same**" means *the atoms that make up its structure,* then *no*
- If "**same**" means *the functionality that it provides,* then *yes*

The wording of the question obscures the very distinctions you most need to center!

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



Are LLMs trustworthy?

What do we mean by trustworthy?

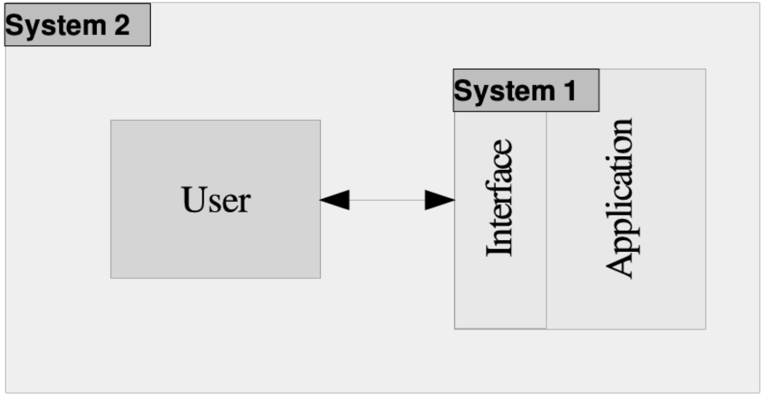
Are they reliable? Are they safe?

What do we mean by reliable? By safe?

- Are you evaluating **"system performance"** at the level of its components (decontextualized)?
- Or **"system performance"** at the level of its *impacts* (contextualized)?

What are we trying to evaluate?

Many LLM Evaluation “riddles” are similarly a definition trick:



System 1: LLM capabilities

What are the quality attributes of this language model?

System 2: Application functionalities

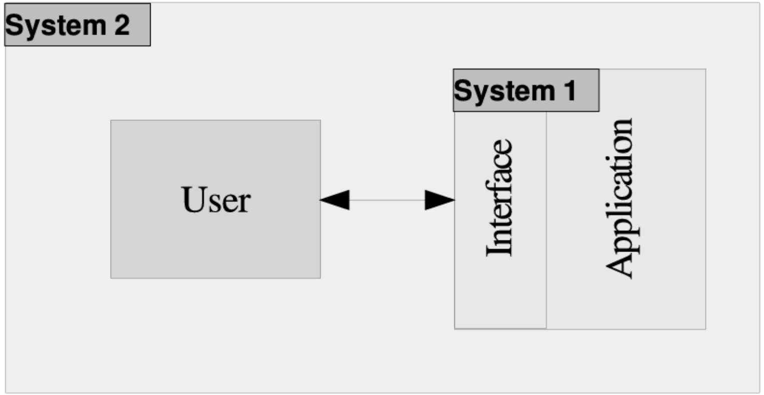
What are the quality attributes of this application

(which is powered, in part, by a language model)?

- Are you evaluating **“system performance”** at the level of its components (decontextualized)?
- Or **“system performance”** at the level of its *impacts* (contextualized)?

What are we trying to evaluate?

Many LLM Evaluation "riddles" are similarly a definition trick:



System 1: LLM capabilities

Processing speed! Rouge score! Token weights!

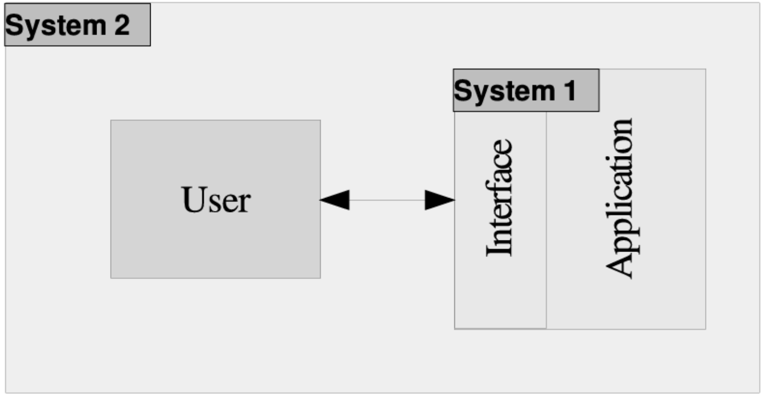
System 2: Application functionalities

Benefits! Risks! Effectiveness! Usefulness! Safety!

- Are you evaluating **"system performance"** at the level of its components (decontextualized)?
- Or **"system performance"** at the level of its *impacts* (contextualized)?

What are we trying to evaluate?

System 2 evaluation is necessary for operational deployment decisions



System 1: LLM capabilities

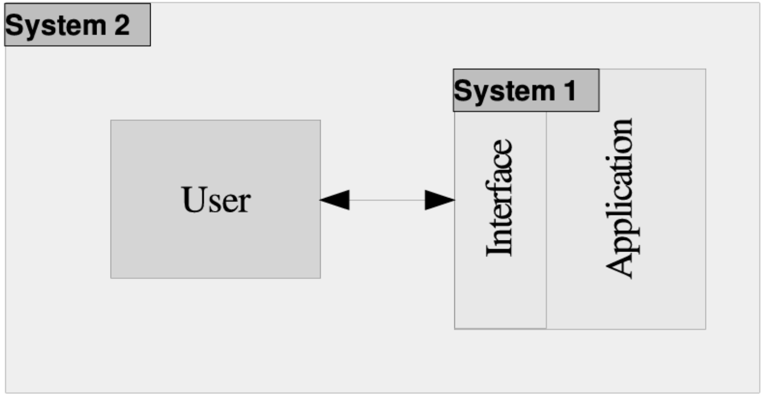
Processing speed! Rouge score! Token weights!

System 2: Application functionalities

Benefits! Risks! Effectiveness! Usefulness! Safety!

What are we trying to evaluate?

System 2 evaluation is necessary for operational deployment decisions



System 1: LLM capabilities

Processing speed! Rouge score! Token weights!

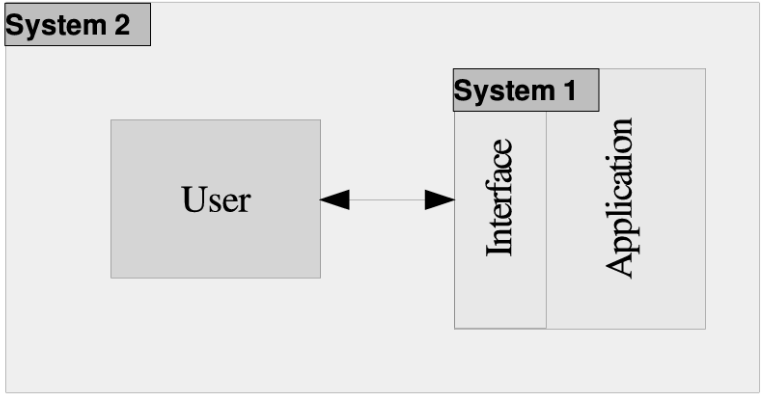
System 2: Application functionalities

Benefits! Risks! Effectiveness! Usefulness! Safety!

The rest of this talk is designed to make it easier for you to "solve the riddle" of what this means for your own deployments

What are we trying to evaluate?

System 2 evaluation is necessary for operational deployment decisions



System 1: LLM capabilities

Processing speed! Rouge score! Token weights!

System 2: Application functionalities

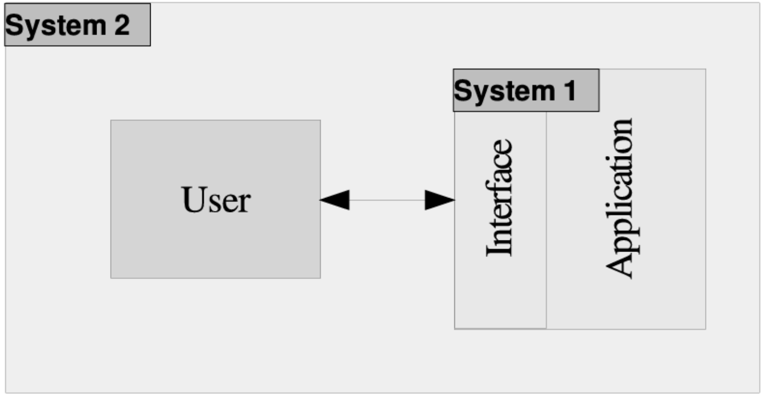
Benefits! Risks! Effectiveness! Usefulness! Safety!

The rest of this talk is designed to make it easier for you to "solve the riddle" of what this means for your own deployments

1. *Define the menu of these functionalities*
2. *Share evaluation considerations for each*

What are we trying to evaluate?

System 2 evaluation is necessary for operational deployment decisions



System 1: LLM capabilities

Processing speed! Rouge score! Token weights!

System 2: Application functionalities

Benefits! Risks! Effectiveness! Usefulness! Safety!

The rest of this talk is designed to make it easier for you to "solve the riddle" of what this means for your own deployments

1. *Define the menu of these functionalities*
2. *Share evaluation considerations for each*
3. *Convince you about discourse analysis ☺*

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.

Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.

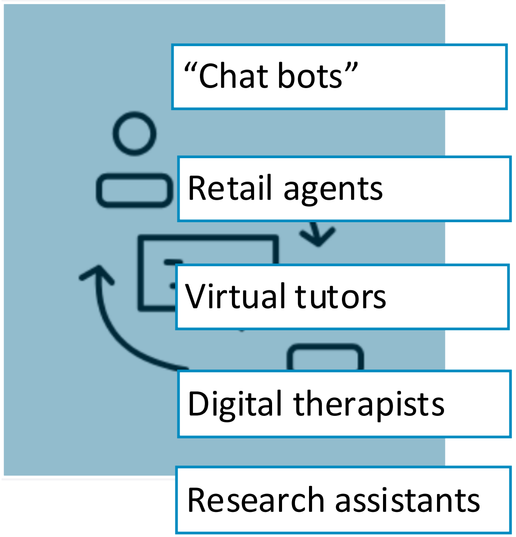
Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.



“Chat bots”

Retail agents

Virtual tutors

Digital therapists

Research assistants

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.

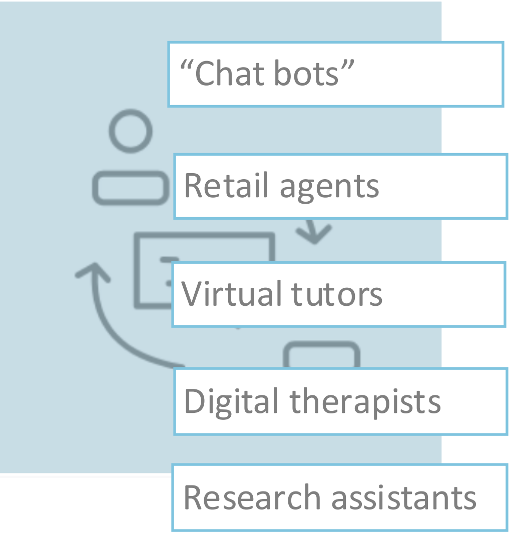
Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.



"Chat bots"

Retail agents

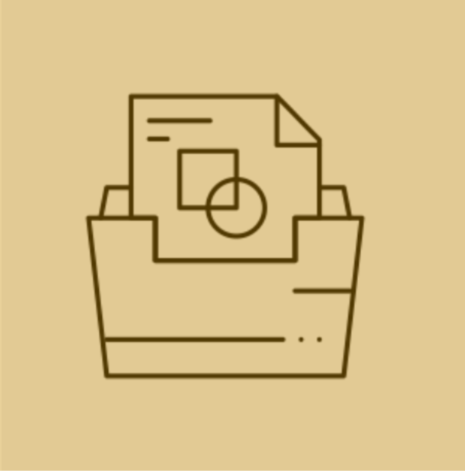
Virtual tutors

Digital therapists

Research assistants

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.



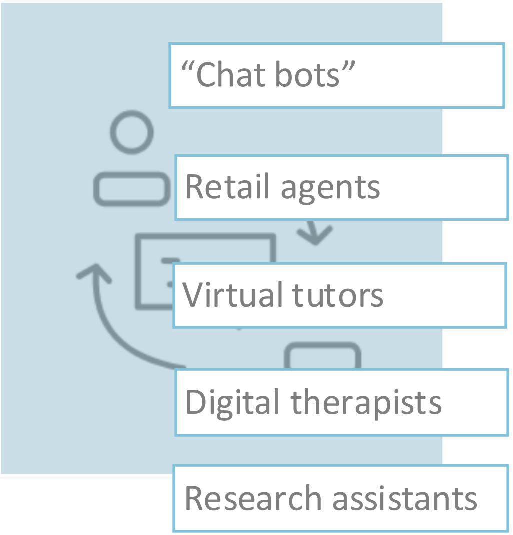
Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.



"Chat bots"

Retail agents

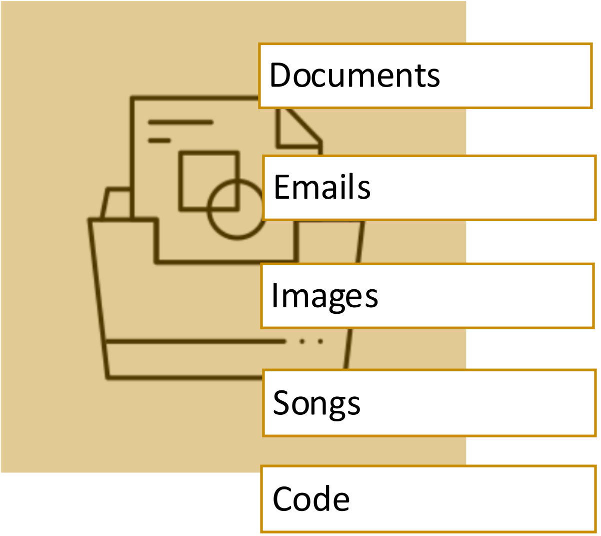
Virtual tutors

Digital therapists

Research assistants

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.



Documents

Emails

Images

Songs

Code

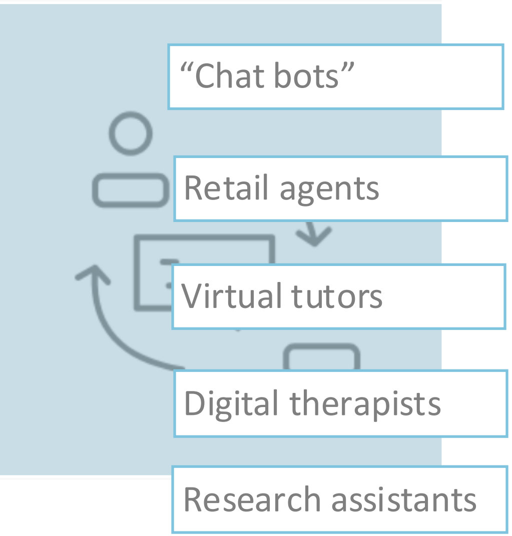
Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.



"Chat bots"

Retail agents

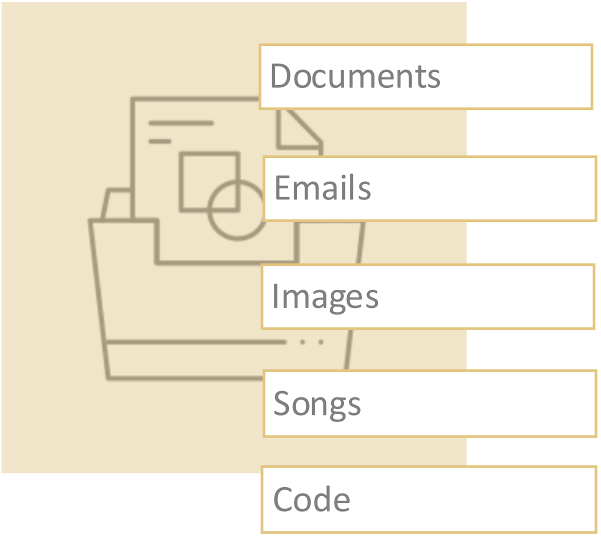
Virtual tutors

Digital therapists

Research assistants

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.



Documents

Emails

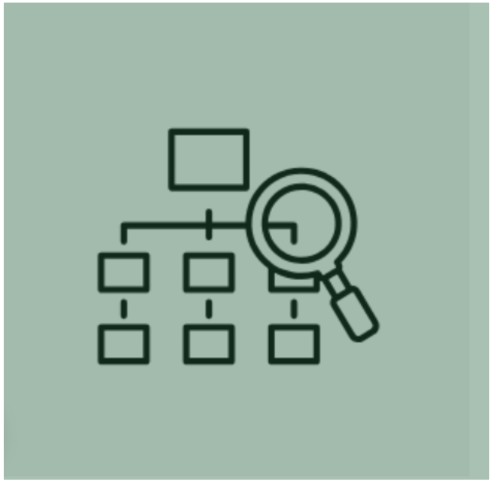
Images

Songs

Code

Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.



LLMs provide 3 categories of application functionality

Conversation

System enables a dialogic interaction where users construct input - and interpret output - through the lens of discourse.

- "Chat bots"
- Retail agents
- Virtual tutors
- Digital therapists
- Research assistants

Generation

System enables user specification of criteria on which to deliver a stand-alone artifact.

- Documents
- Emails
- Images
- Songs
- Code

Analysis

System enables transformation of language signals into different signals, in accordance with identified specifications.

- Grammar checking
- Theme-finding
- Translation
- Threat detection
- Agentic decisioning

Let’s apply this to an example: intelligence analysis

Conversation

Intelligence ideation

The analyst and LLM discuss intel across sources to strengthen the interpretive scope and rigor.

Generation

Document summarization

The analyst feeds in a long intelligence report (or set of reports), and the LLM generates a summary that retains the “most important” information, in accordance with specifications.

Analysis

Propaganda detection

The system ingests streams of data (e.g., sourced reports, messages, news articles) and uses an LLM to examine those documents for signals of potential adversarial propaganda or influence.

Let's apply this to an example: intelligence analysis

Conversation

Intelligence ideation

The analyst and LLM discuss intel across sources to strengthen the interpretive scope and rigor.

The LLM can support structure and ideation across considerations like:

- **Brainstorming** *(identify and interrogate alternative explanations)*
- **Surfacing ambiguities** *(identifying blind spots, open questions, testing edge-cases, poking at assumptions)*
- **Contextualizing** *(applying intel across situations to reveal subtle patterns or applications)*

Generation

Document summarization

The analyst feeds in a long intelligence report (or set of reports), and the LLM generates a summary that retains the "most important" information, in accordance with specifications.

Analysis

Propaganda detection

The system ingests streams of data (e.g., sourced reports, messages, news articles) and uses an LLM to examine those documents for signals of potential adversarial propaganda or influence.

Let's apply this to an example: intelligence analysis

Conversation

Intelligence ideation

The analyst and LLM discuss intel across sources to strengthen the interpretive scope and rigor.

The LLM can support structure and ideation across considerations like:

- **Brainstorming** *(identify and interrogate alternative explanations)*
- **Surfacing ambiguities** *(identifying blind spots, open questions, testing edge-cases, poking at assumptions)*
- **Contextualizing** *(applying intel across situations to reveal subtle patterns or applications)*

Generation

Document summarization

The analyst feeds in a long intelligence report (or set of reports), and the LLM generates a summary that retains the "most important" information, in accordance with specifications.

Those specifications can span:

- **importance** *(scrutiny / prioritization appropriate across detail types)*
- **phrasing** *(the level of paraphrasing allowable / desirable)*
- **formatting** *(output requirements)*

Analysis

Propaganda detection

The system ingests streams of data (e.g., sourced reports, messages, news articles) and uses an LLM to examine those documents for signals of potential adversarial propaganda or influence.

Let's apply this to an example: intelligence analysis

Conversation

Intelligence ideation

The analyst and LLM discuss intel across sources to strengthen the interpretive scope and rigor.

The LLM can support structure and ideation across considerations like:

- **Brainstorming** *(identify and interrogate alternative explanations)*
- **Surfacing ambiguities** *(identifying blind spots, open questions, testing edge-cases, poking at assumptions)*
- **Contextualizing** *(applying intel across situations to reveal subtle patterns or applications)*

Generation

Document summarization

The analyst feeds in a long intelligence report (or set of reports), and the LLM generates a summary that retains the "most important" information, in accordance with specifications.

Those specifications can span:

- **importance** *(scrutiny / prioritization appropriate across detail types)*
- **phrasing** *(the level of paraphrasing allowable / desirable)*
- **formatting** *(output requirements)*

Analysis

Propaganda detection

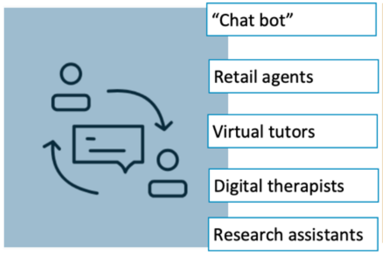
The system ingests streams of data (e.g., sourced reports, messages, news articles) and uses an LLM to examine those documents for signals of potential adversarial propaganda or influence.

System 1: signal detection (*how the LLM is trained to detect adversarial signals)*

System 2: signal response *(how the system triggers actions following detection)*

What does this mean for evaluation?

Conversation



- "Chat bot"
- Retail agents
- Virtual tutors
- Digital therapists
- Research assistants

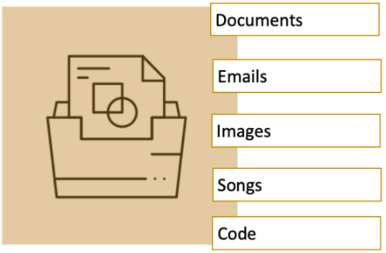
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – "co-constructed meaning" with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



- Documents
- Emails
- Images
- Songs
- Code

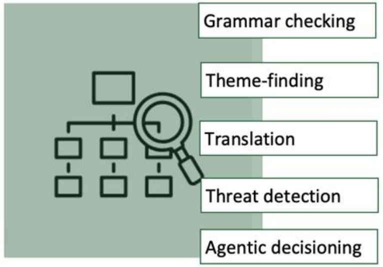
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



- Grammar checking
- Theme-finding
- Translation
- Threat detection
- Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

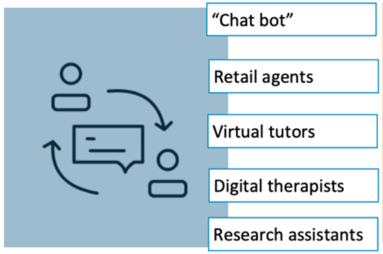
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy,* tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness,* determined by system design decisions.

What does this mean for evaluation?

Conversation



“Chat bot”

Retail agents

Virtual tutors

Digital therapists

Research assistants

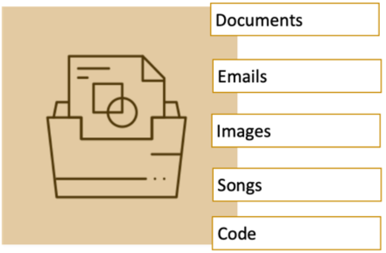
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – “co-constructed meaning” with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



Documents

Emails

Images

Songs

Code

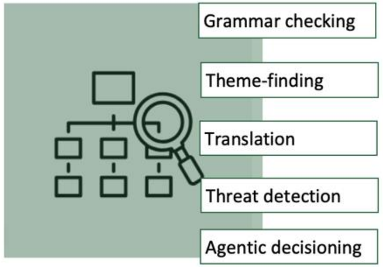
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



Grammar checking

Theme-finding

Translation

Threat detection

Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

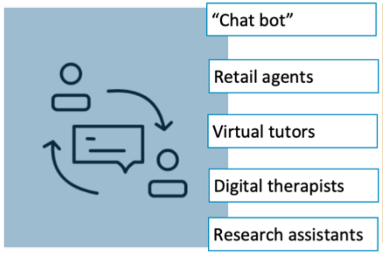
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy*, tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness*, determined by system design decisions.

What does this mean for evaluation?

Conversation


"Chat bot"
Retail agents
Virtual tutors
Digital therapists
Research assistants

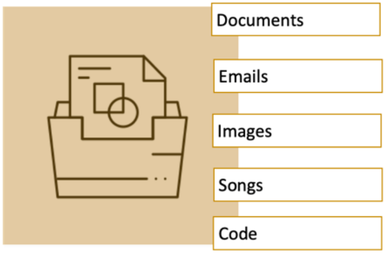
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – "co-constructed meaning" with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation


Documents
Emails
Images
Songs
Code

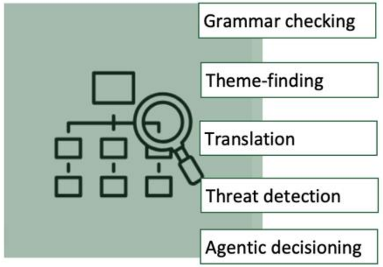
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis


Grammar checking
Theme-finding
Translation
Threat detection
Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

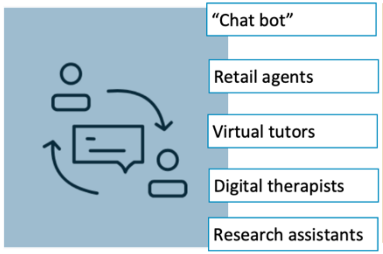
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy*, tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness*, determined by system design decisions.

What does this mean for evaluation?

Conversation



- "Chat bot"
- Retail agents
- Virtual tutors
- Digital therapists
- Research assistants

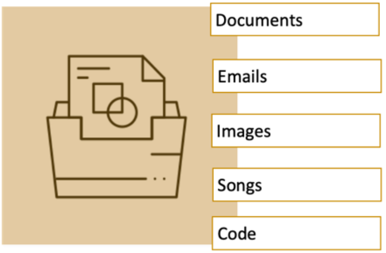
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – "co-constructed meaning" with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



- Documents
- Emails
- Images
- Songs
- Code

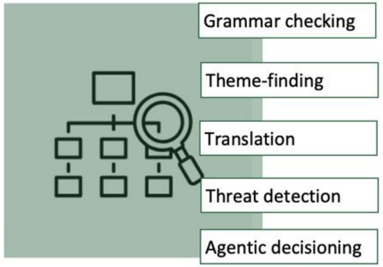
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



- Grammar checking
- Theme-finding
- Translation
- Threat detection
- Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

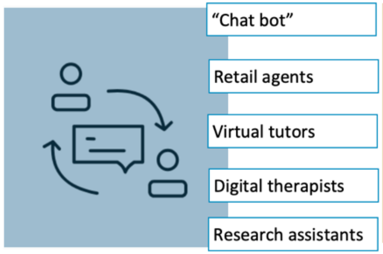
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy,* tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness,* determined by system design decisions.

What does this mean for evaluation?

Conversation



- "Chat bot"
- Retail agents
- Virtual tutors
- Digital therapists
- Research assistants

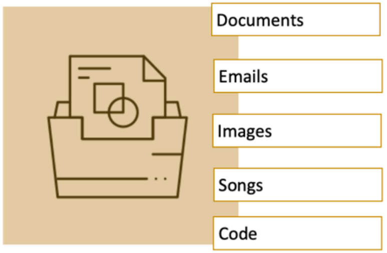
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – "co-constructed meaning" with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



- Documents
- Emails
- Images
- Songs
- Code

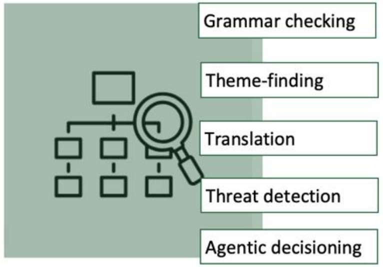
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



- Grammar checking
- Theme-finding
- Translation
- Threat detection
- Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

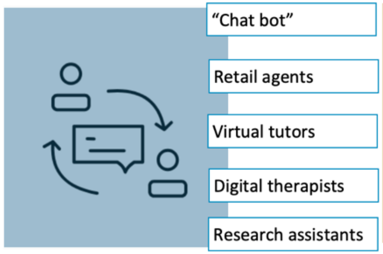
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy*, tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness*, determined by system design decisions.

What does this mean for evaluation?

Conversation



“Chat bot”

Retail agents

Virtual tutors

Digital therapists

Research assistants

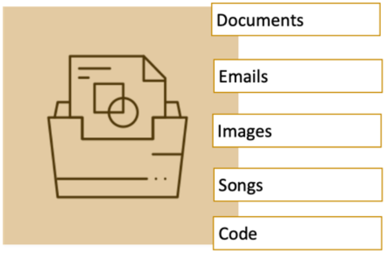
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – “co-constructed meaning” with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



Documents

Emails

Images

Songs

Code

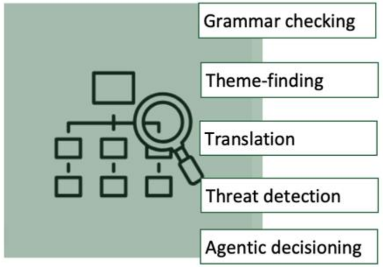
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



Grammar checking

Theme-finding

Translation

Threat detection

Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

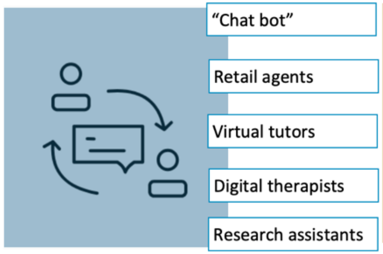
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy*, tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness*, determined by system design decisions.

Focus: *evaluating conversation*

Conversation



"Chat bot"

Retail agents

Virtual tutors

Digital therapists

Research assistants

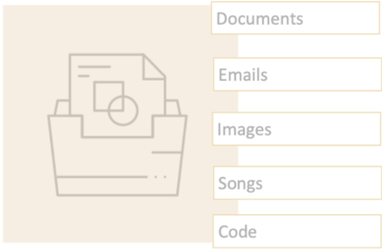
The *thing* is the interaction.

Success = the quality of the discourse.

Control panel = pragmatic *fluency* – "co-constructed meaning" with LLM.

Meaningful evaluation requires metapragmatic considerations across the discourse frame. *(aka context)*

Generation



Documents

Emails

Images

Songs

Code

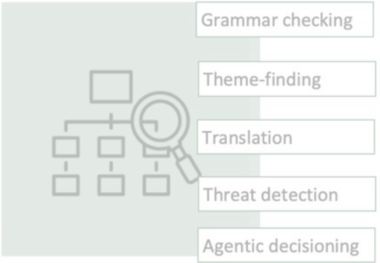
The *thing* is the artifact.

Success = the quality of the delivery.

Control panel = the UI to input criteria. It may or may not involve language.

Meaningful evaluation must center success criteria defined at the *artifact* level *(like standard HCI eval.)*

Analysis



Grammar checking

Theme-finding

Translation

Threat detection

Agentic decisioning

The *thing* is the signal…
… but really, what you *do* with that signal.

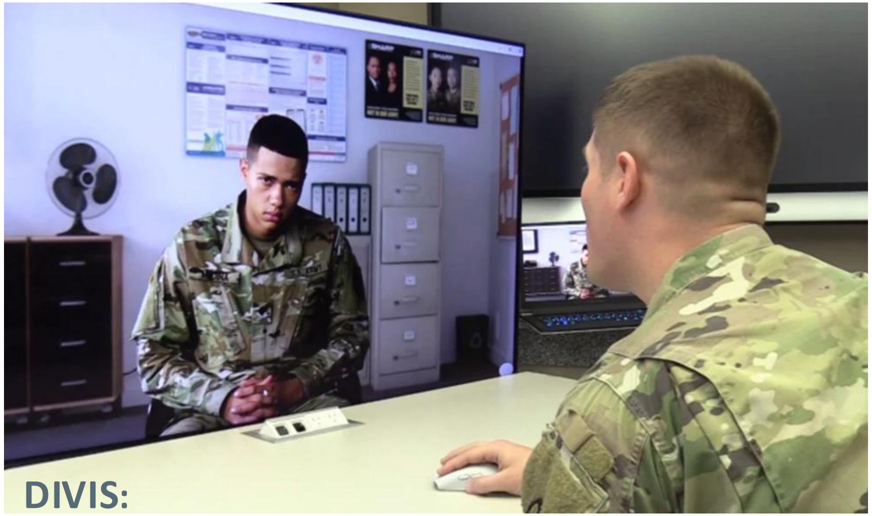
Meaningful evaluation *especially* requires distinguishing *accuracy* from *impacts.*

System 1 (signal-as-detection): *Accuracy*, tuned by training data, criteria, thresholds.

System 2 (signal-as-trigger): *Appropriateness*, determined by system design decisions.

Focus: *evaluating conversation*

Evaluating conversational *functionalities* requires applying conversational *methodologies*.



DIVIS:

Goal: provide environment for victim advocate students to practice leading highly-emotional sexual assault intake interviews

- First **conversational agent** in the 1960s: ELIZA
- Has since been applied to dozens of different DoD applications
- People apply or adapt their *human-human* language norms to *human-agent* language experiences: **useful for evaluation!**

PAL3:

Goal: *on-the-job training*

Battle buddy:

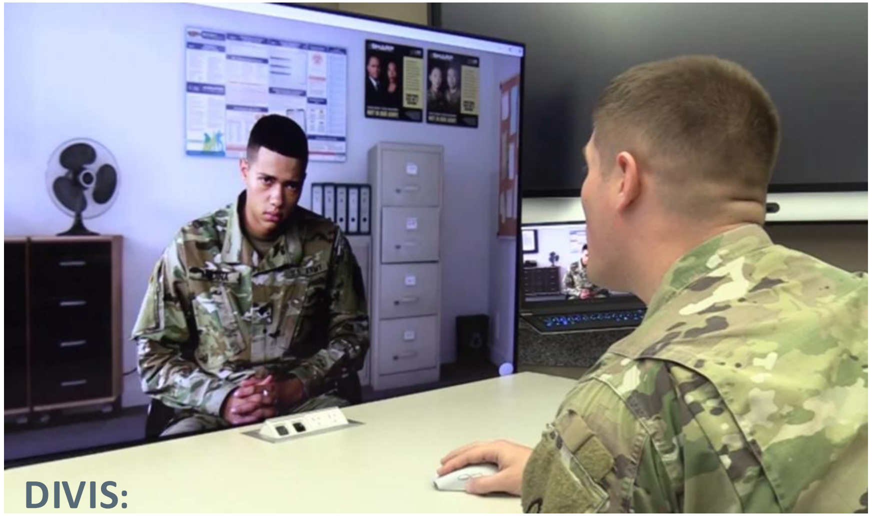
Goal: *veteran life quality*

VITA4Vets:

Goal: *interviewing skills*

Focus: *evaluating conversation*

Evaluating conversational *functionalities* requires applying conversational *methodologies*.



DIVIS:

Goal: provide environment for victim advocate students to practice leading highly-emotional sexual assault intake interviews

- **Discourse analysis:** how to evaluate language in context
  - **Meaning is constructed across multiple turns.**
    How can you tell if "nice job" sincere praise or sarcastic indictment?
  - **Roles are explicitly and implicitly negotiated.**
    Who am I in this conversation? Who are you?
    What type of conversation are we having?
  - **Communication success requires:**
    Theory of mind: What does this person know?

    Grounding & Repair: Given that, what should I say?
    Did they know what I mean? How can I get us on the same page?

Focus: *evaluating conversation*

Evaluating conversational *functionalities* requires applying conversational *methodologies*.

This might initially sound like complications…

One-shot I/O metrics frequently not realistic

Success metrics depend on the *type of discourse*

Different levels of generative flexibility appropriate for brainstorming vs document summarization.

Prioritize "disambiguation" over "accuracy"

- **Discourse analysis:** how to evaluate language in context
  - **Meaning is constructed across multiple turns.**
    How can you tell if "nice job" sincere praise or sarcastic indictment?
  - **Roles are explicitly and implicitly negotiated.**
    Who am I in this conversation? Who are you?
    What type of conversation are we having?
  - **Communication success requires:**
    Theory of mind: What does this person know?

    Grounding & Repair: Given that, what should I say?
    Did they know what I mean? How can I get us on the same page?

Focus: *evaluating conversation*

Evaluating conversational *functionalities* requires applying conversational *methodologies*.

This might initially sound like complications…

One-shot I/O metrics frequently not realistic

Success metrics depend on the *type of discourse*

Different levels of generative flexibility appropriate for brainstorming vs document summarization.

Prioritize ”disambiguation” over “accuracy”

…except that we have 60+ years of work to pull from!

- **Discourse analysis:** how to evaluate language in context
  - **Meaning is constructed across multiple turns.**
    How can you tell if “nice job” sincere praise or sarcastic indictment?
  - **Roles are explicitly and implicitly negotiated.**
    Who am I in this conversation? Who are you?
    What type of conversation are we having?
  - **Communication success requires:**
    Theory of mind: What does this person know?

    Grounding & Repair: Given that, what should I say?
    Did they know what I mean? How can I get us on the same page?

Focus: *evaluating conversation*

Evaluating conversational *functionalities* requires applying conversational *methodologies*.

This might initially sound like complications…

One-shot I/O metrics frequently not realistic

Success metrics depend on the *type of discourse*

Different levels of generative flexibility appropriate for brainstorming vs document summarization.

Prioritize ”disambiguation” over “accuracy”

…except that we have 60+ years of work to pull from!

- **Discourse analysis:** how to evaluate language in context
  - **Meaning is constructed across multiple turns.**
    How can you tell if "nice job" sincere praise or sarcastic indictment?
  - **Roles are explicitly and implicitly negotiated.**
    Who am I in this conversation? Who are you?
    What type of conversation are we having?
  - **Communication success requires:**
    Theory of mind: What does this person know?

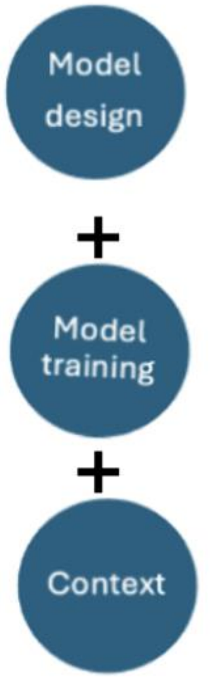
    Grounding & Repair: Given that, what should I say?
    Did they know what I mean? How can I get us on the same page?

1. Discourse is explainable

2. Dialogue is designable

Discourse is explainable

LLM performance requires task alignment across three pillars:



Capabilities:
What the model can do

Information:
What the model knows

Interpretation:
What the model “thinks its doing”

Discourse is explainable

LLM performance requires task alignment across three pillars:

Model design
+
Model training
+
Context

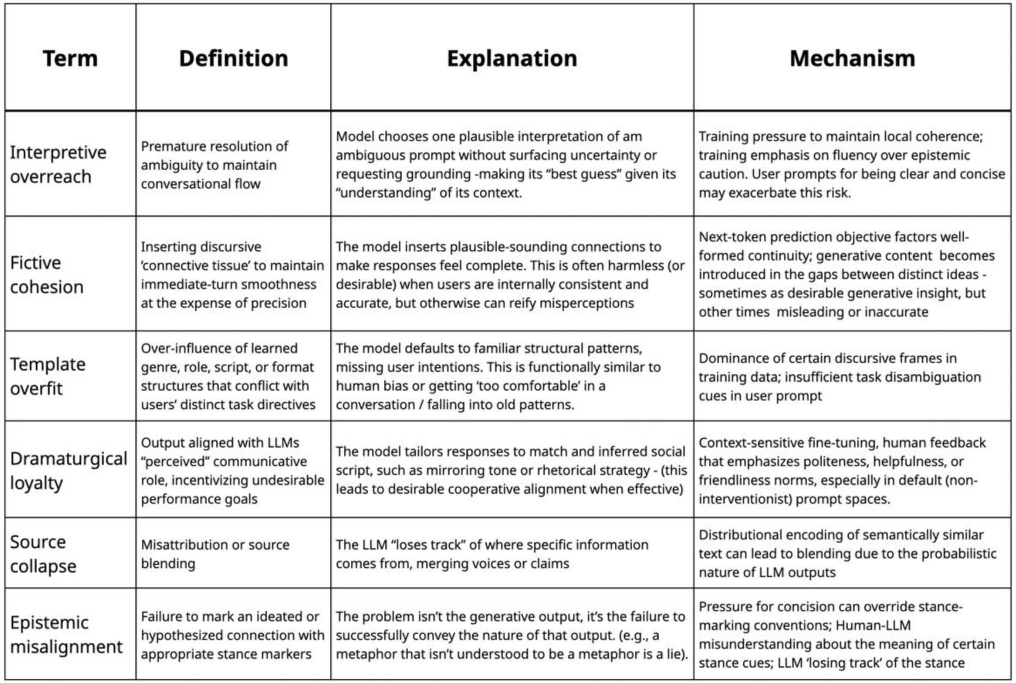
Capabilities:
What the model can do

Information:
What the model knows

Interpretation:
What the model "thinks its doing"

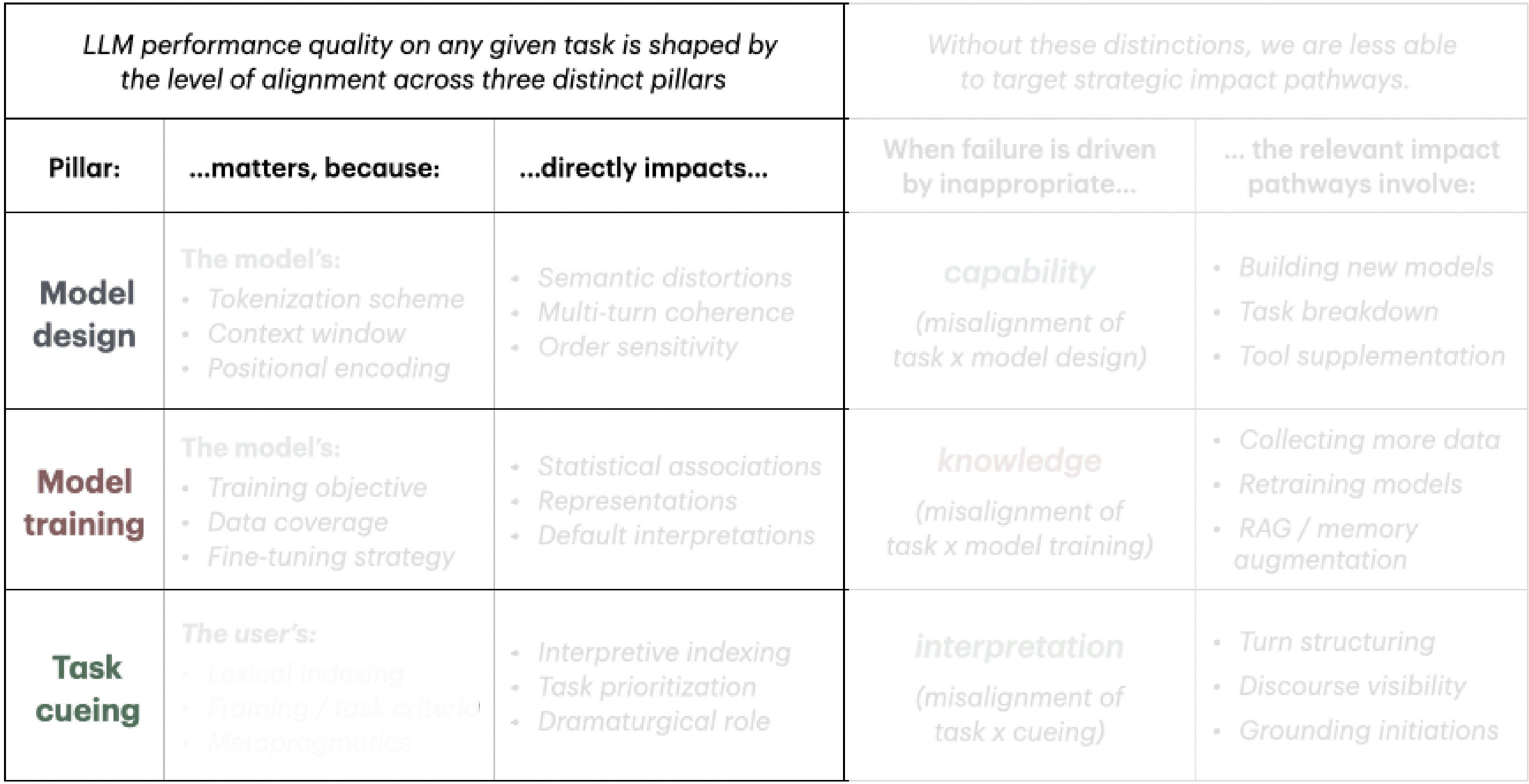
Each of these pillars leads to different *types of problems.*

Right now, we call them all hallucinations.



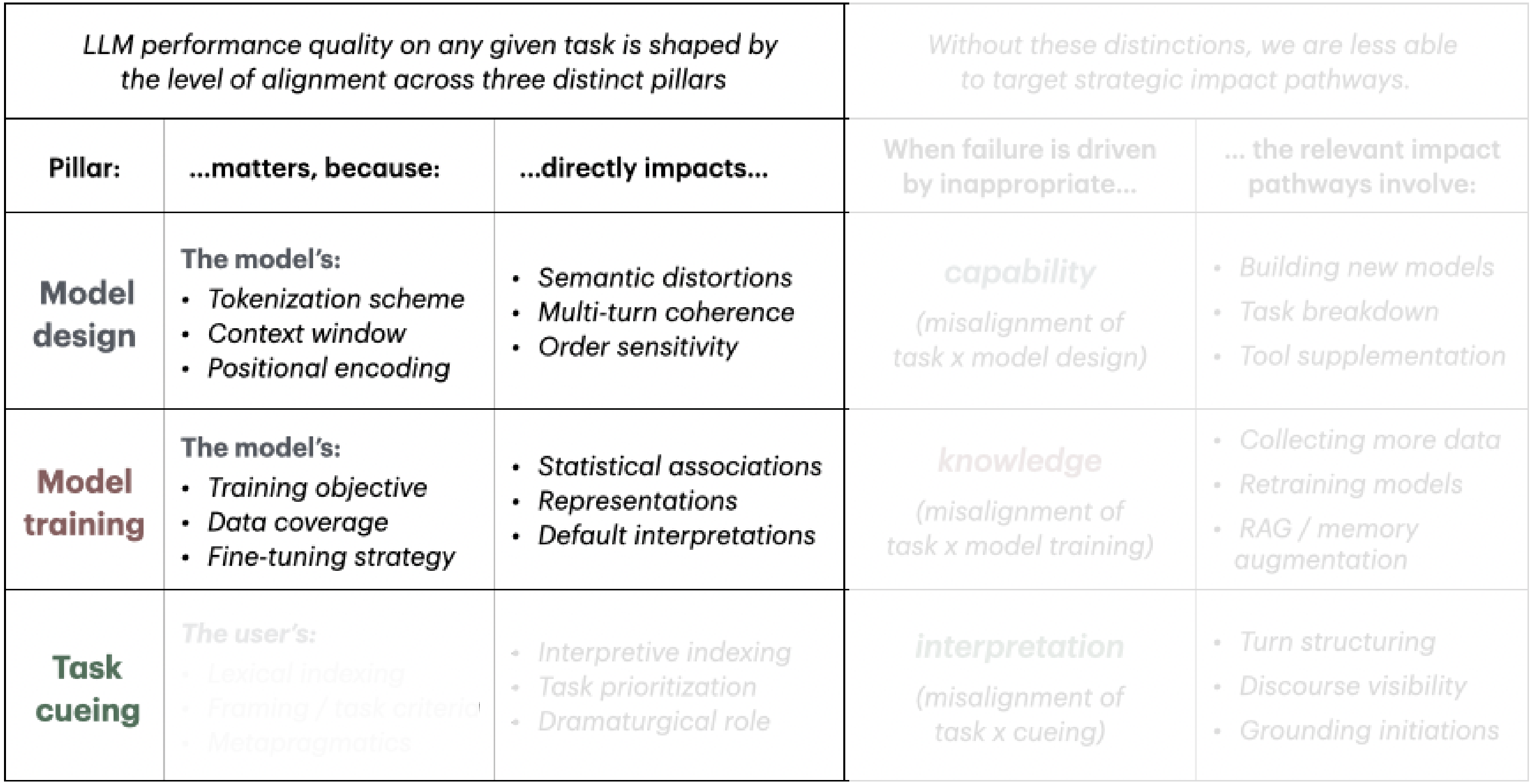
Term	Definition	Explanation	Mechanism
Interpretive overreach	Premature resolution of ambiguity to maintain conversational flow	Model chooses one plausible interpretation of am ambiguous prompt without surfacing uncertainty or requesting grounding -making its "best guess" given its "understanding" of its context.	Training pressure to maintain local coherence; training emphasis on fluency over epistemic caution. User prompts for being clear and concise may exacerbate this risk.
Fictive cohesion	Inserting discursive 'connective tissue' to maintain immediate-turn smoothness at the expense of precision	The model inserts plausible-sounding connections to make responses feel complete. This is often harmless (or desirable) when users are internally consistent and accurate, but otherwise can reify misperceptions	Next-token prediction objective factors well-formed continuity; generative content becomes introduced in the gaps between distinct ideas - sometimes as desirable generative insight, but other times misleading or inaccurate
Template overfit	Over-influence of learned genre, role, script, or format structures that conflict with users' distinct task directives	The model defaults to familiar structural patterns, missing user intentions. This is functionally similar to human bias or getting 'too comfortable' in a conversation / falling into old patterns.	Dominance of certain discursive frames in training data; insufficient task disambiguation cues in user prompt
Dramaturgical loyalty	Output aligned with LLMs "perceived" communicative role, incentivizing undesirable performance goals	The model tailors responses to match and inferred social script, such as mirroring tone or rhetorical strategy - (this leads to desirable cooperative alignment when effective)	Context-sensitive fine-tuning, human feedback that emphasizes politeness, helpfulness, or friendliness norms, especially in default (non-interventionist) prompt spaces.
Source collapse	Misattribution or source blending	The LLM "loses track" of where specific information comes from, merging voices or claims	Distributional encoding of semantically similar text can lead to blending due to the probabilistic nature of LLM outputs
Epistemic misalignment	Failure to mark an ideated or hypothesized connection with appropriate stance markers	The problem isn't the generative output, it's the failure to successfully convey the nature of that output. (e.g., a metaphor that isn't understood to be a metaphor is a lie).	Pressure for concision can override stance-marking conventions; Human-LLM misunderstanding about the meaning of certain stance cues; LLM 'losing track' of the stance

Dialogue is designable



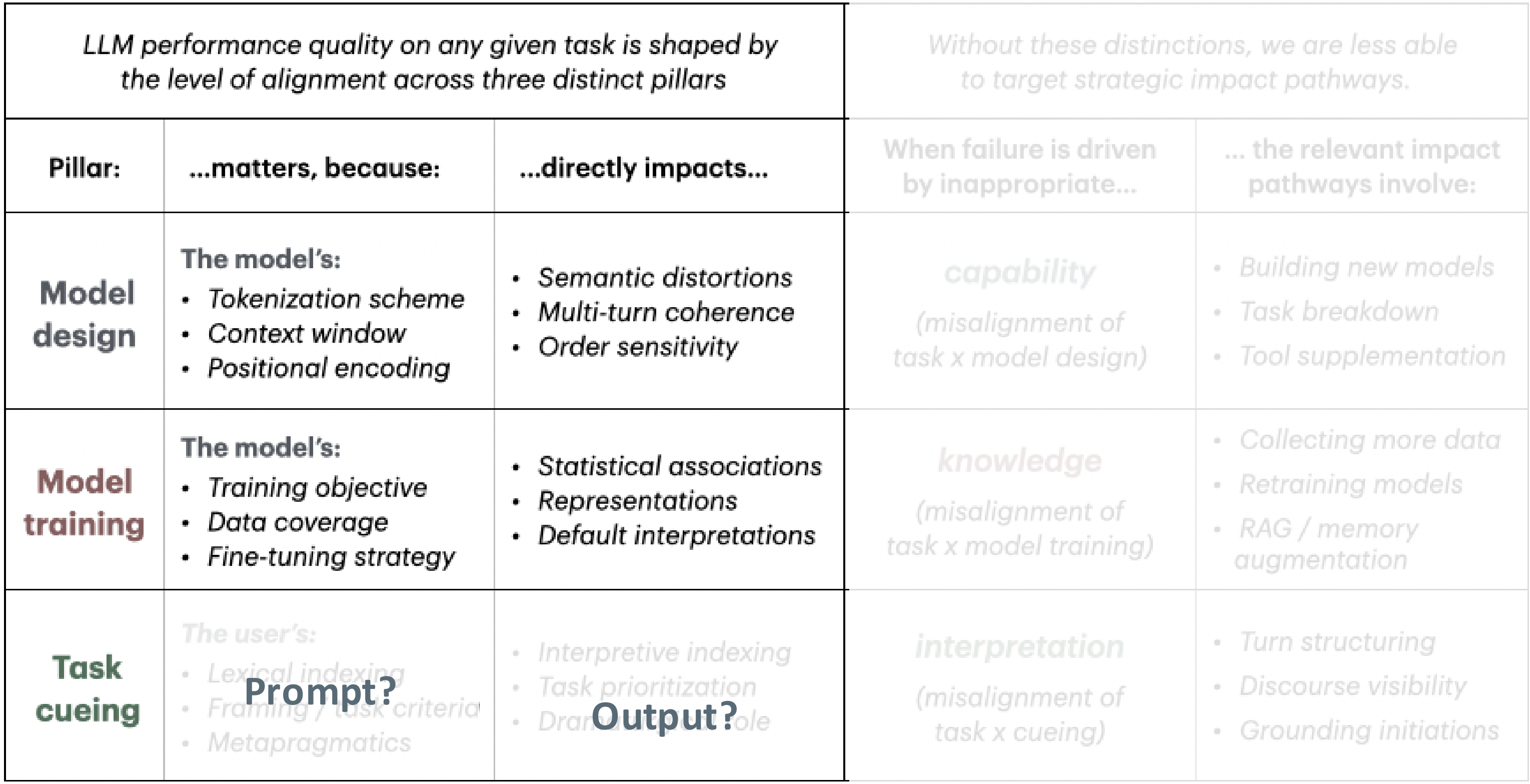
LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	**The user's:** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	***interpretation*** *(misalignment of task x cueing)*	• *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Dialogue is designable



LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	***The user's:*** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	***interpretation*** *(misalignment of task x cueing)*	• *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Dialogue is designable

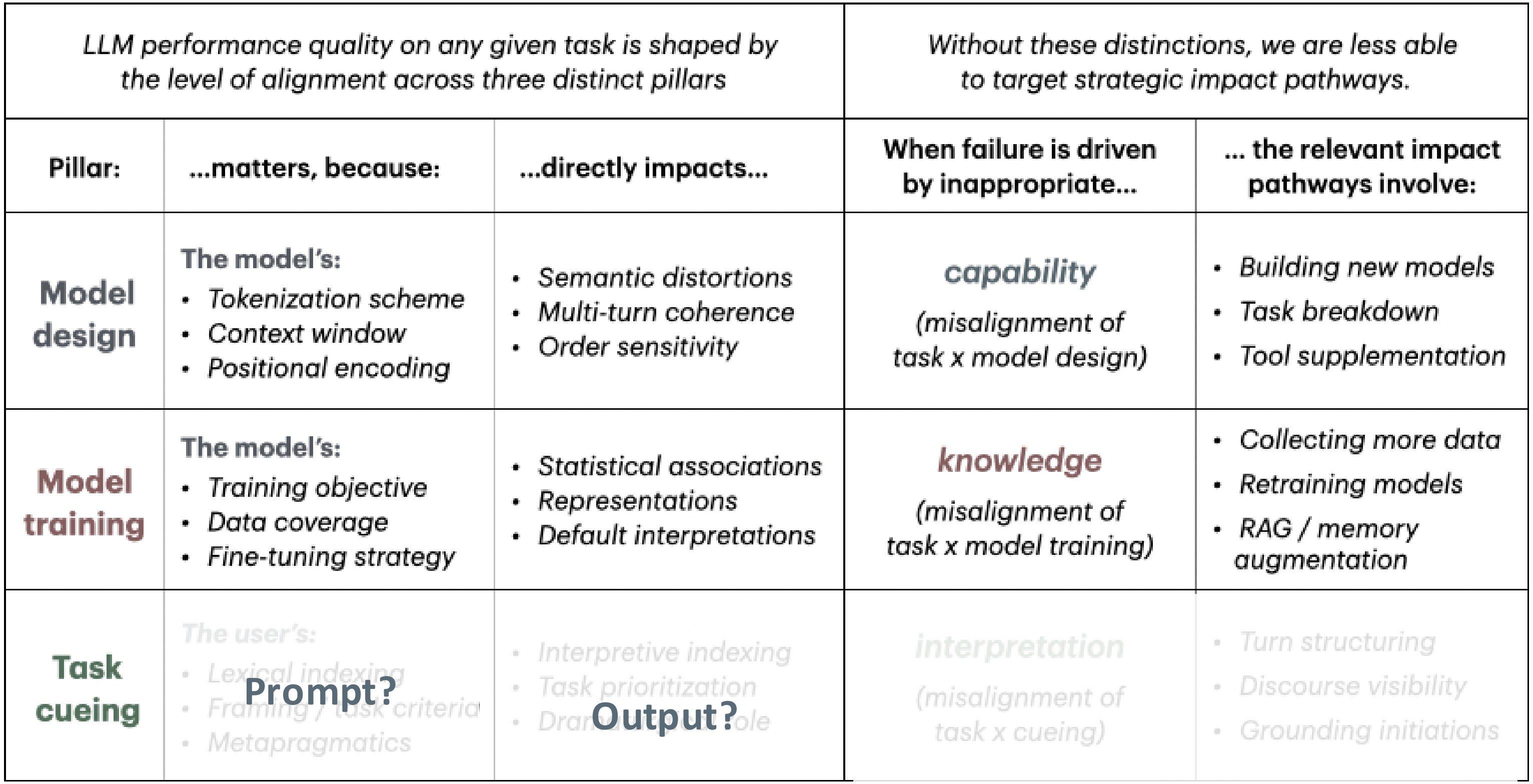


LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			Without these distinctions, we are less able to target strategic impact pathways.	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	**The user's:** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	***interpretation*** *(misalignment of task x cueing)*	• *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Prompt?

Output?

Dialogue is designable

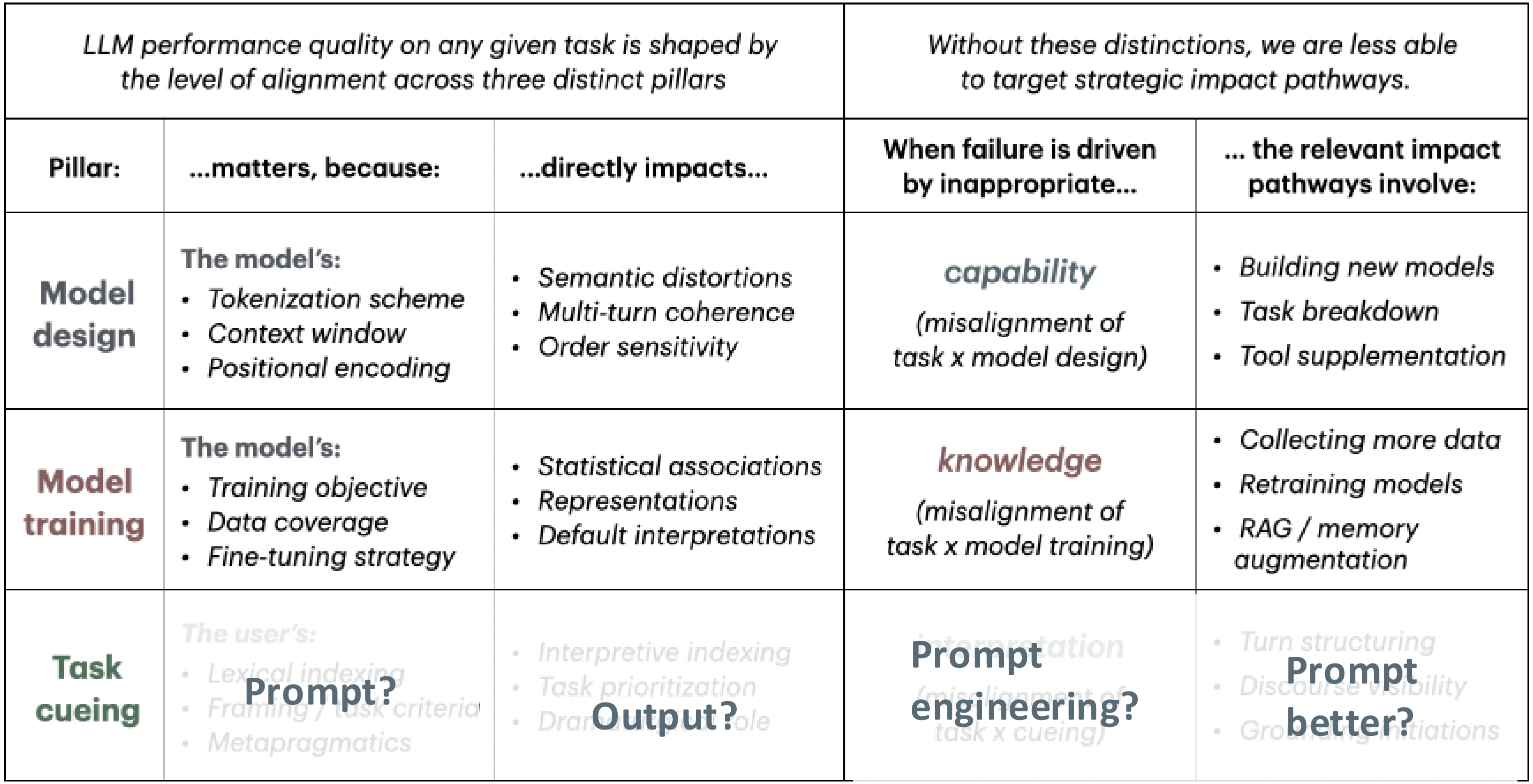


LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	**The user's:** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	***interpretation*** *(misalignment of task x cueing)*	• *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Prompt?

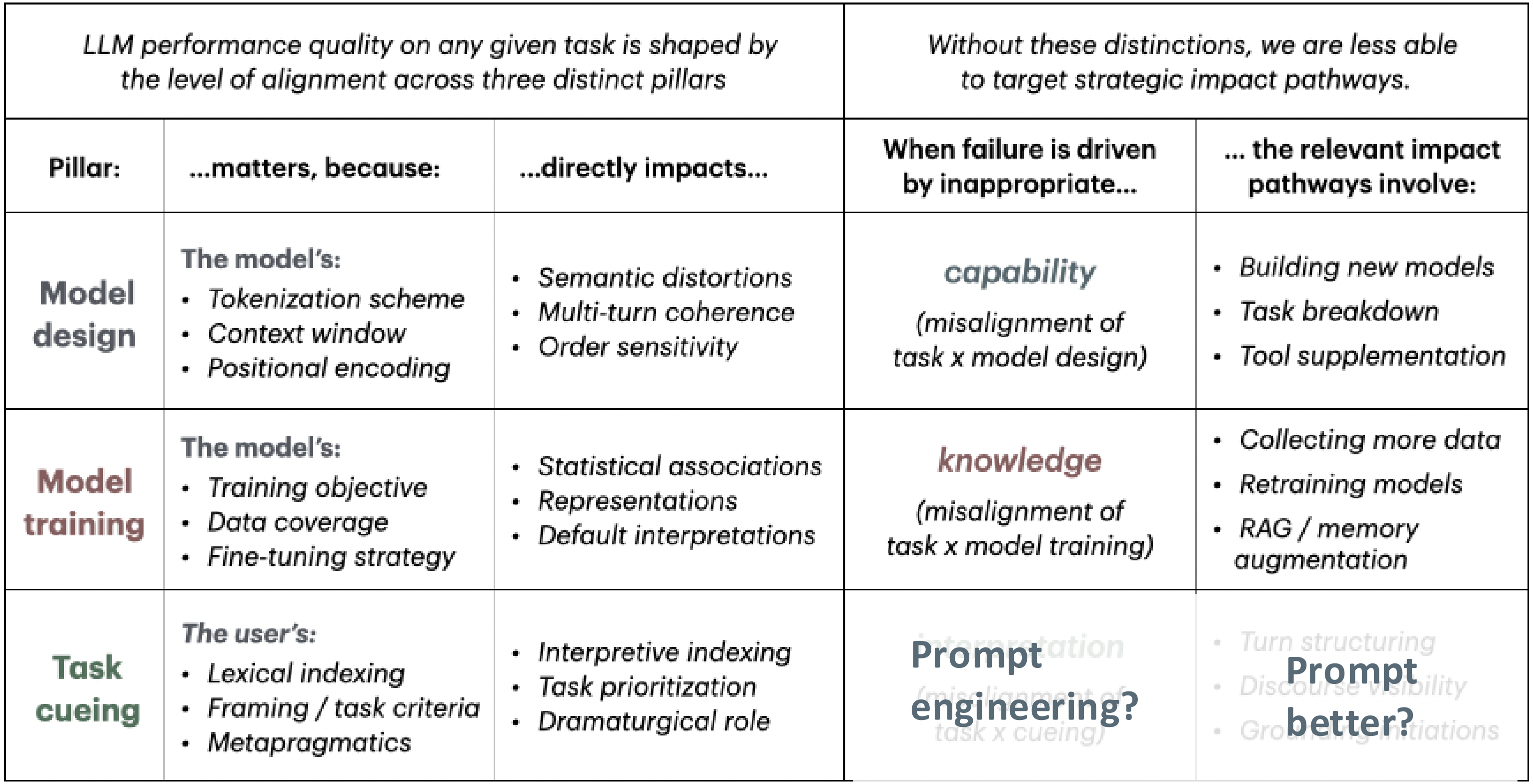
Output?

Dialogue is designable



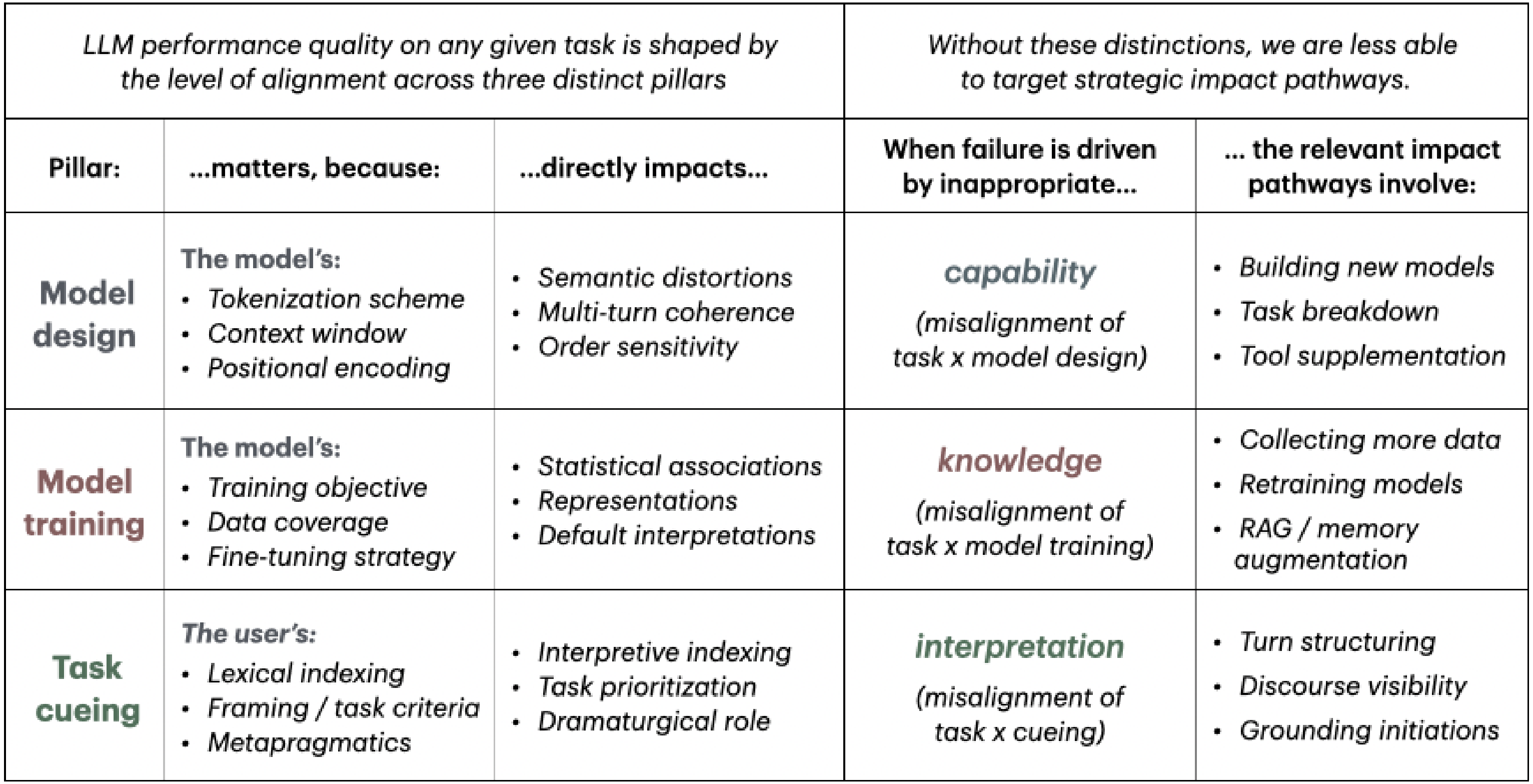
LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	**The user's:** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics* **Prompt?**	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role* **Output?**	***interpretation*** *(misalignment of task x cueing)* **Prompt engineering?**	• *Turn structuring* • *Discourse visibility* • *Grounding initiations* **Prompt better?**

Dialogue is designable



LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	***The user's:*** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	**Prompt engineering?** ***interpretation*** *(misalignment of task x cueing)*	**Prompt better?** • *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Dialogue is designable



LLM performance quality on any given task is shaped by the level of alignment across three distinct pillars			*Without these distinctions, we are less able to target strategic impact pathways.*	
Pillar:	**...matters, because:**	**...directly impacts...**	**When failure is driven by inappropriate...**	**... the relevant impact pathways involve:**
Model design	**The model's:** • *Tokenization scheme* • *Context window* • *Positional encoding*	• *Semantic distortions* • *Multi-turn coherence* • *Order sensitivity*	***capability*** *(misalignment of task x model design)*	• *Building new models* • *Task breakdown* • *Tool supplementation*
Model training	**The model's:** • *Training objective* • *Data coverage* • *Fine-tuning strategy*	• *Statistical associations* • *Representations* • *Default interpretations*	***knowledge*** *(misalignment of task x model training)*	• *Collecting more data* • *Retraining models* • *RAG / memory augmentation*
Task cueing	***The user's:*** • *Lexical indexing* • *Framing / task criteria* • *Metapragmatics*	• *Interpretive indexing* • *Task prioritization* • *Dramaturgical role*	***interpretation*** *(misalignment of task x cueing)*	• *Turn structuring* • *Discourse visibility* • *Grounding initiations*

Discourse analysis as a diagnostic lens

1. **Identify** what type of discourse failure occurred

- *Surface where dialogue breaks down (e.g., interpretive overreach, failed implicature…)*
- *Treat chat logs as structured evidence – rich dataset that you have for free!*

2. **Explain** the mechanism behind the misalignment

- *Surface potential hypotheses that explain patterns from established literature*
- *Experiments: contrastive trials contrasting sociolinguistic explanations*
- *Center explanatory mechanism in the science of evaluation*

3. **Intervene** at the level of the interaction

- *Design and test discourse-level fixes (scaffolds, stance markers, role visibility).*
- *Evaluate not just system accuracy, but alignment and coordination gains.*