New Dimensions of Evaluation Amid the Rise of Large Language Models

Enrique Amigó

Associate Professor

Universidad Nacional de Educación a Distancia

TIGER Speaker Bio

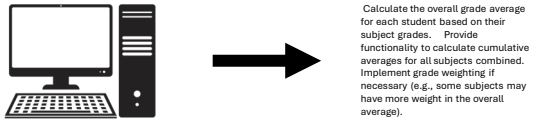
- **Position & Affiliation**: Associate Professor at UNED
- **International Evaluation Campaigns**: WEPS-3, Replab 2012, Replab 2013, Detox 2020, and Exists 2023
- **Research Focus**: System evaluation, text representation, evaluation metrics and text representation spaces
- **Publications & Impact**: Over 3,000 citations on Google Scholar
- **Conference Service**: Served as General Chair for SIGIR 2022
- **Honors & Awards**:
  - The SEPLN National Award for Best Research Monograph in Natural Language Processing
  - The Google Faculty Research Award in 2012
- **Project Experience**: European and national projects, including initiatives to measure the technology gap between Spanish and English in language technologies
- **Current Project**: The ODESIA contract project funded by the Ministry of Economic Affairs and Digital Transformation, focusing on developing metrics and indicators to identify the gap between English and Spanish technologies

Labelling based tasks

(Classification, Ranking, Clustering, Information Extraction…)

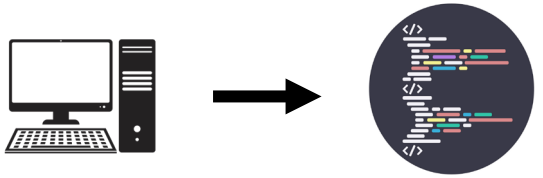
Natural Language Generation

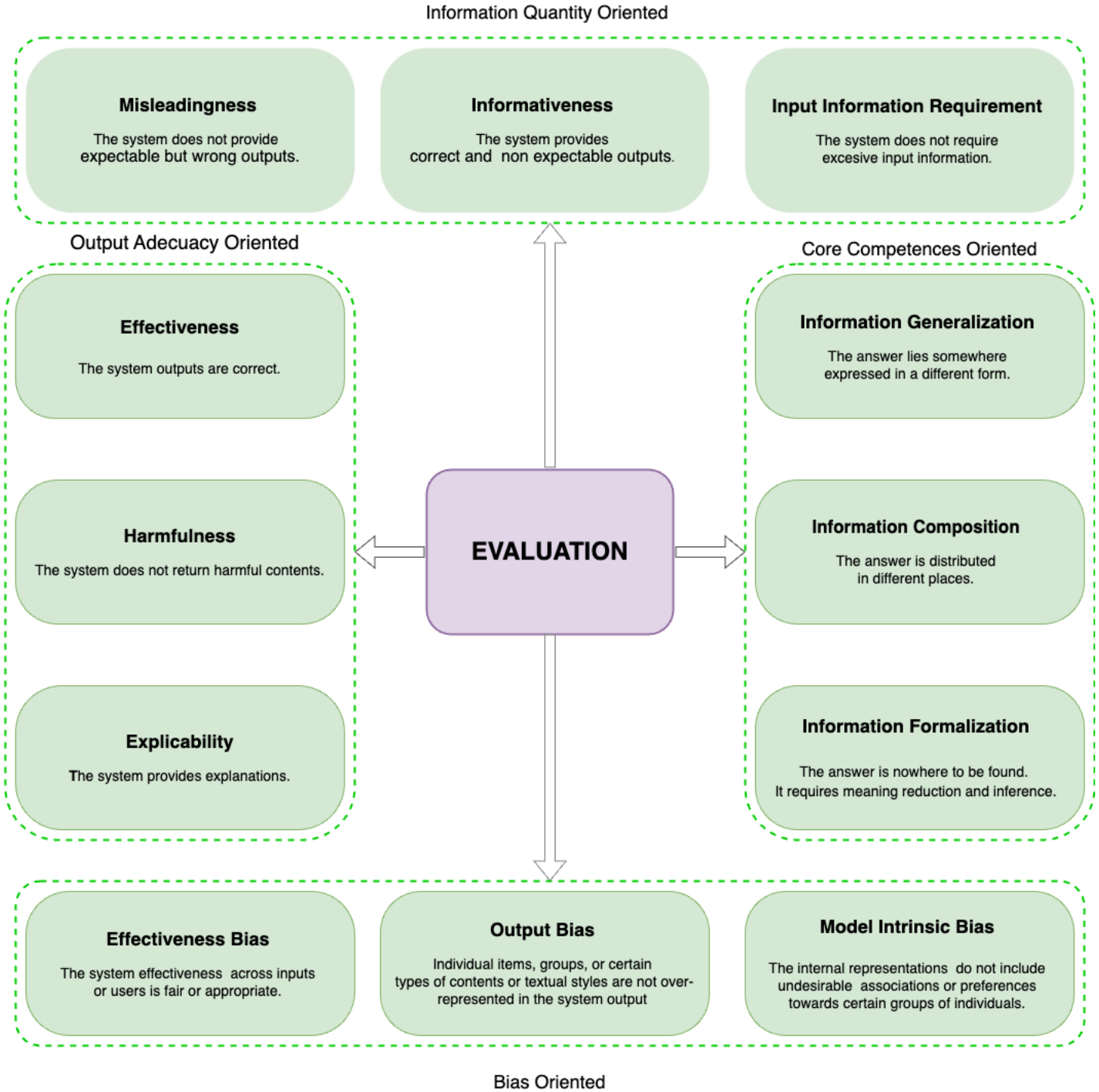
(Virtual assistants, Summarization, Q&A,…)

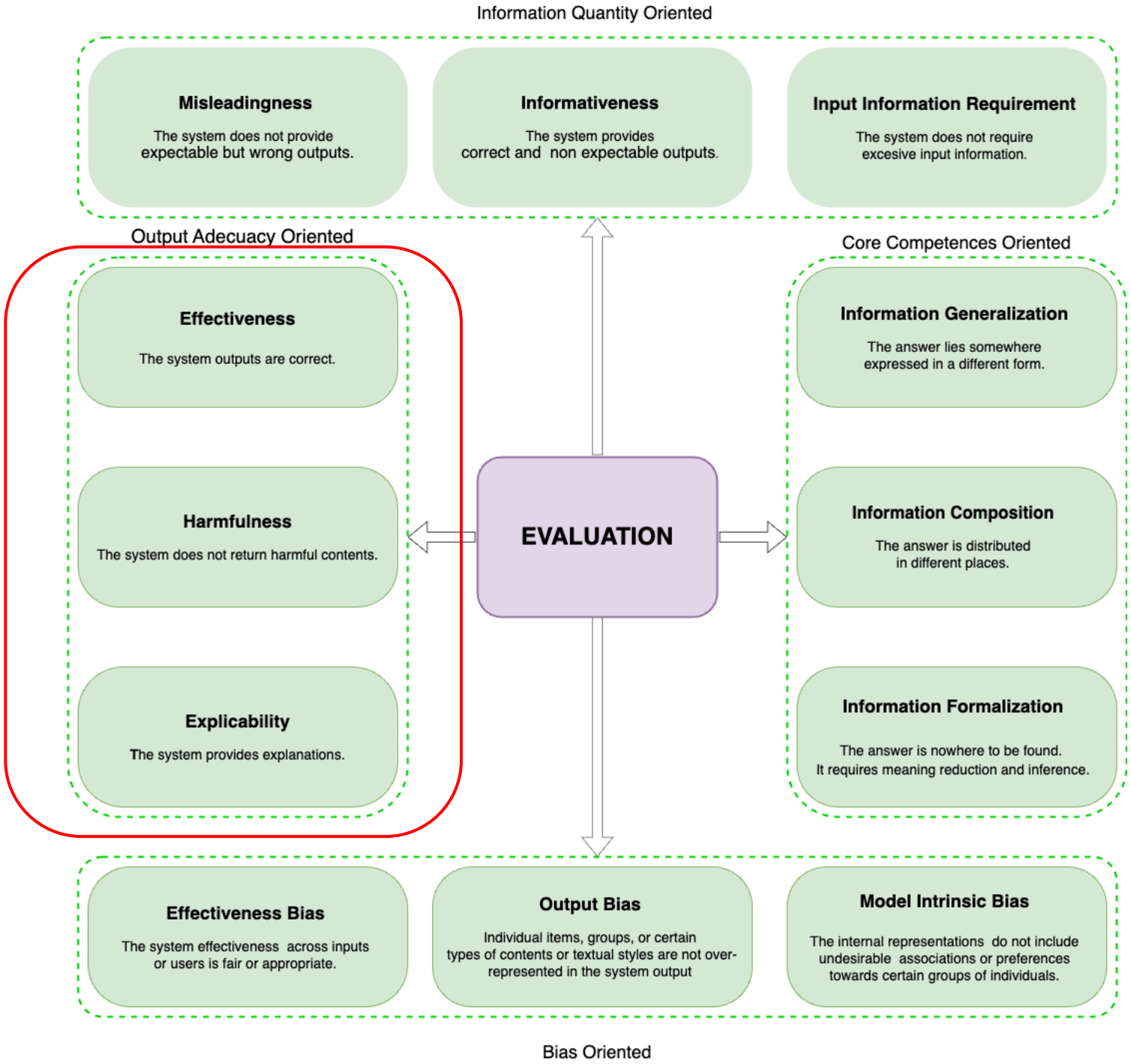


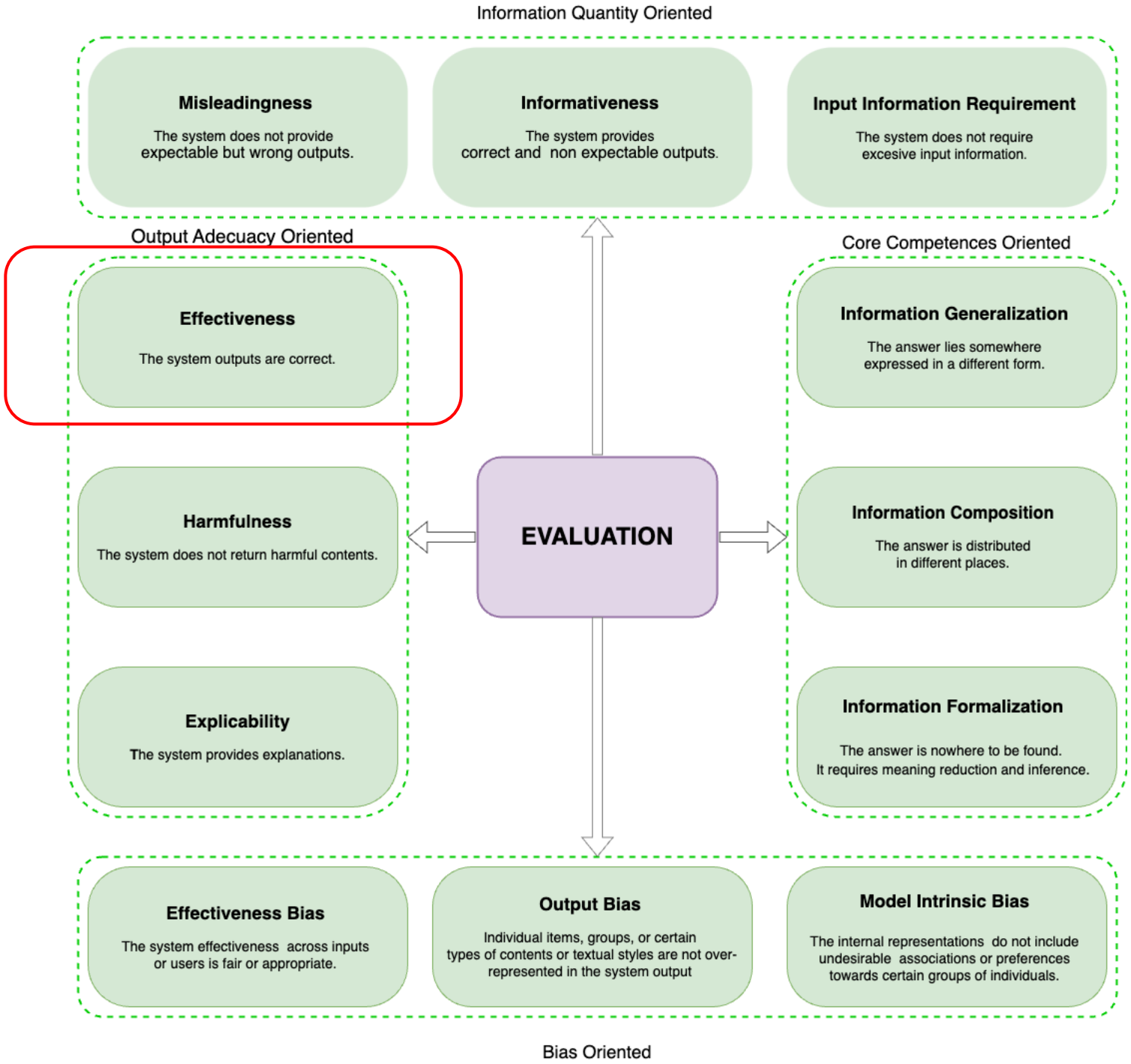
Formal Language Generation

(text to SQL, Code Generation, Semantic Parsing)



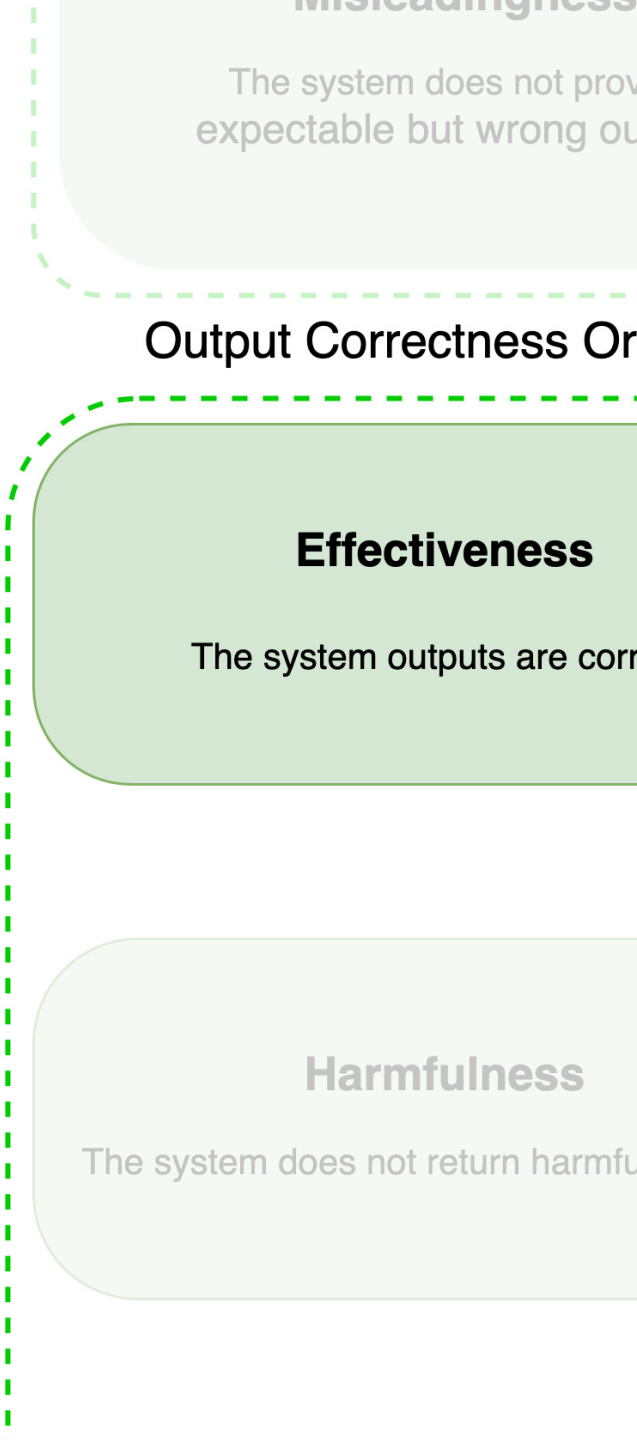

Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented


Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented


Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

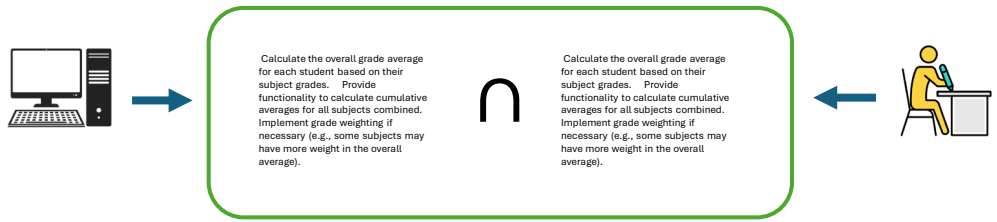
- Labelling based tasks
  - Simple (Accuracy, F measure, nDCG, matching based and counting based clustering metrics),
  - Sophisticated: Information theory based: Classification (Amigo and Delgado, 2022) clustering (Meila, 2007), information retrieval (Amigó et al., 2022). User models in IR (Chapelle et al., 2009; Moffat and Zobel, 2008). Mixed tasks: Diversification (Amigó et al., 2018b), hierarchical classification (Amigo and Delgado, 2022), disagreement (Basile et al., 2021).

- Formal Language Generation (Code, SQL, Semantic Parsing)
  - Token overlap: BLEU (Python (Austin et al., 2021b), SQL (Yu et al., 2019)) Lacks in semantics.
  - Exact matching: (Kim and Linzen, 2020; Li et al., 2021a, Yu et al., 2018). Restricted to simple outputs.
  - Semantic equivalence: (Python code tests (Chen et al., 2021; Liang et al., 2023; Kulal et al., 2019). False positives.



- Text generation:
  - Word overlap: ROUGE (Lin, 2004) BLEU (Papineni et al., 2002), ,CIDER, NIST, GTM, HLEPOR, RIBES, MASI, WER, TER, DICE

    Lacks in language variation.



Misleadingness

The system does not prov
expectable but wrong ou

Output Correctness Or

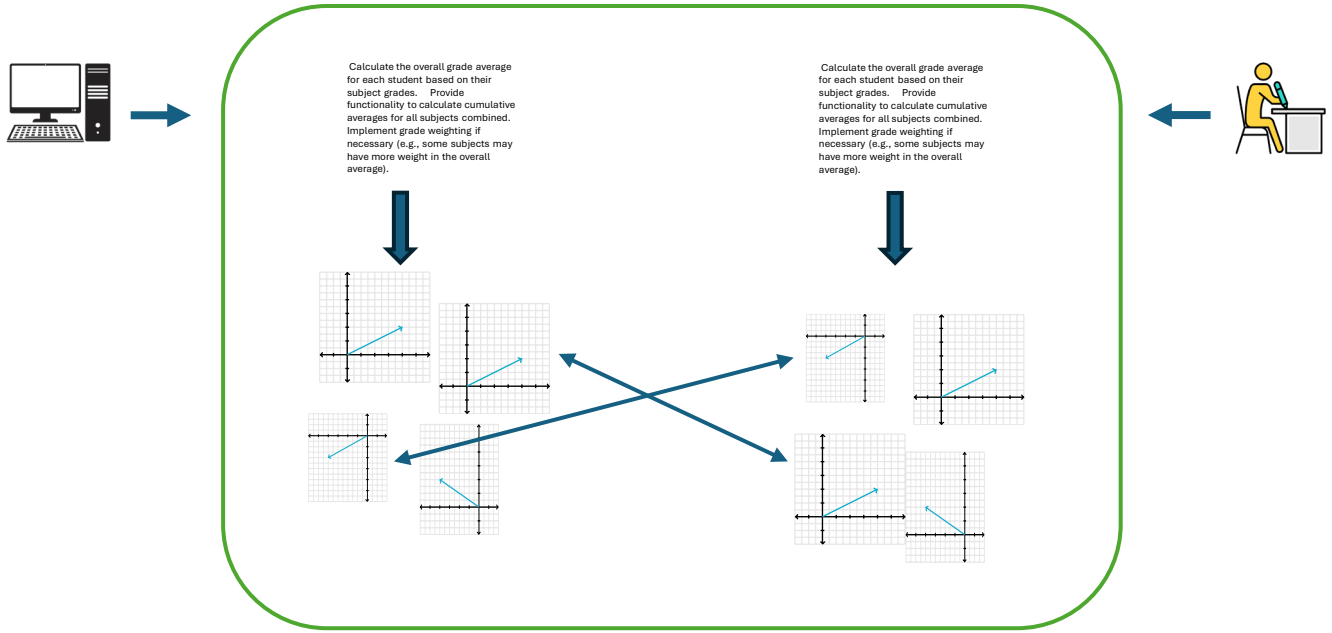
Effectiveness

The system outputs are corr

Harmfulness

The system does not return harmfu

- Text generation:
  - Word overlap: ROUGE (Lin, 2004) BLEU (Papineni et al., 2002), ,CIDER, NIST, GTM, HLEPOR, RIBES, MASI, WER, TER, DICE
    Lacks in language variation.
  - Lexical embedding alignment: MAUVE, BertScore (Zhang et al., 2020), MAUVE, MEANT 2.0, YISI, WMD o SMD.
    Distributional semantics bias



The system does not prov
expectable but wrong ou

Output Correctness Or

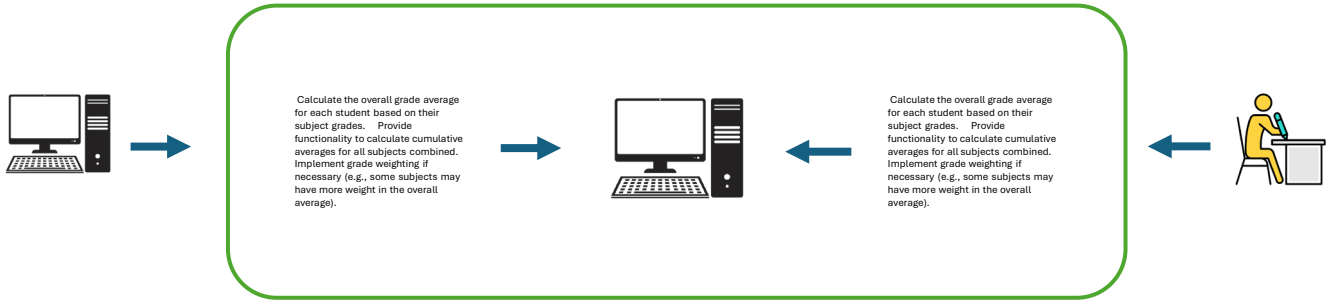
Effectiveness

The system outputs are corr

Harmfulness

The system does not return harmfu

- Text generation:
  - Word overlap: ROUGE (Lin, 2004) BLEU (Papineni et al., 2002), ,CIDER, NIST, GTM, HLEPOR, RIBES, MASI, WER, TER, DICE Lacks in language variation.
  - Lexical embedding alignment: MAUVE, BertScore (Zhang et al., 2020), MAUVE, MEANT 2.0, YISI, WMD o SMD. Distributional semantics bias
  - Language processing tools: Textual similarity, textual entailment, natural language inference (Celikyilmaz et al., 2020) NLP tool bias
  - Trained models: Finetuning BARTScore (Yuan et al., 2021), Specific quality aspects (Kryscinski et al., 2020a; Wang et al., 2020; Cao et al., 2020) Training bias
  - Prompt based: (Liu et al. 2023, Fu et al. 2024; Yuan, Neubig, and Liu 2021) LLM bias.



The system does not prov
expectable but wrong ou

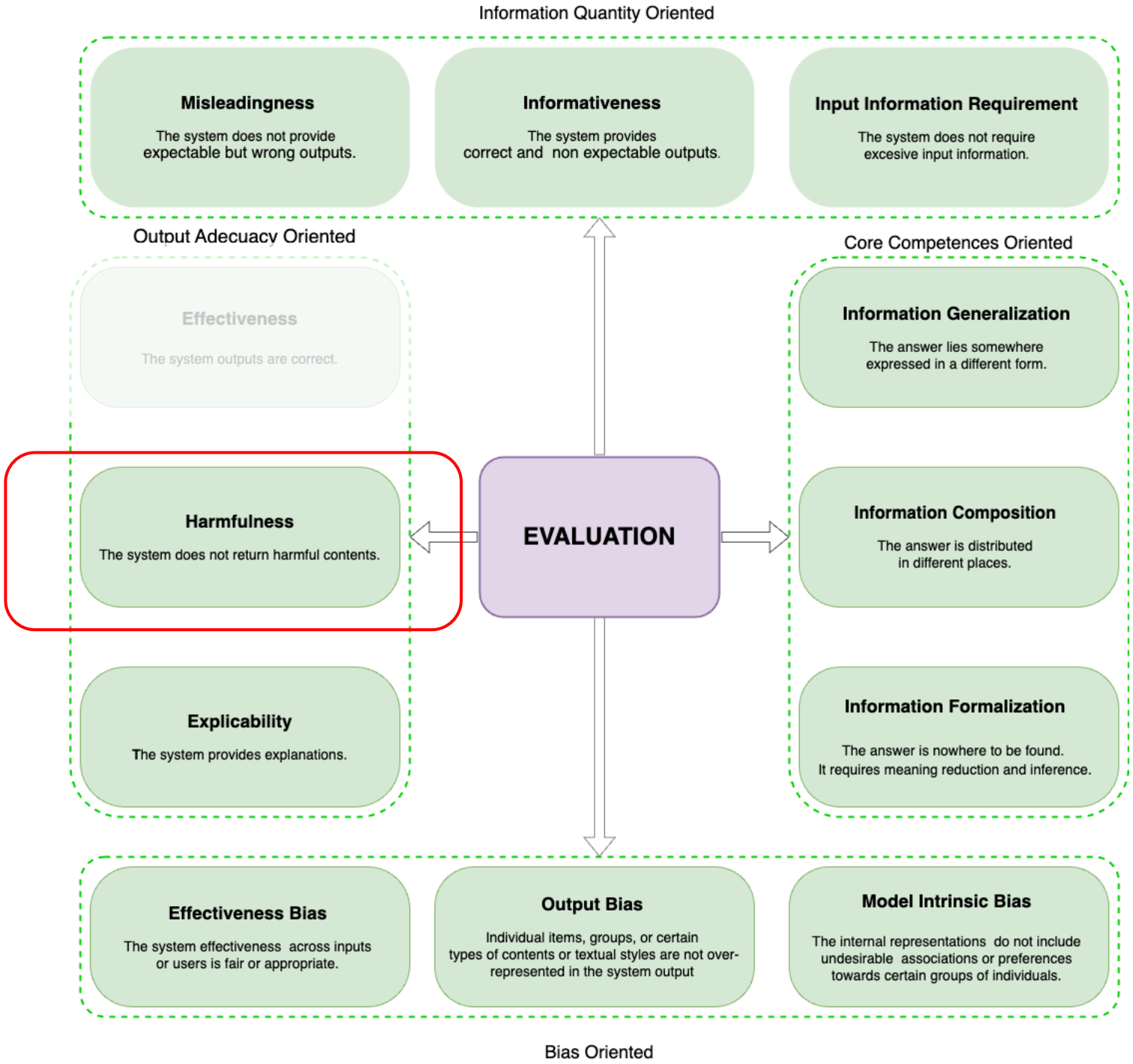
Output Correctness Or

Effectiveness

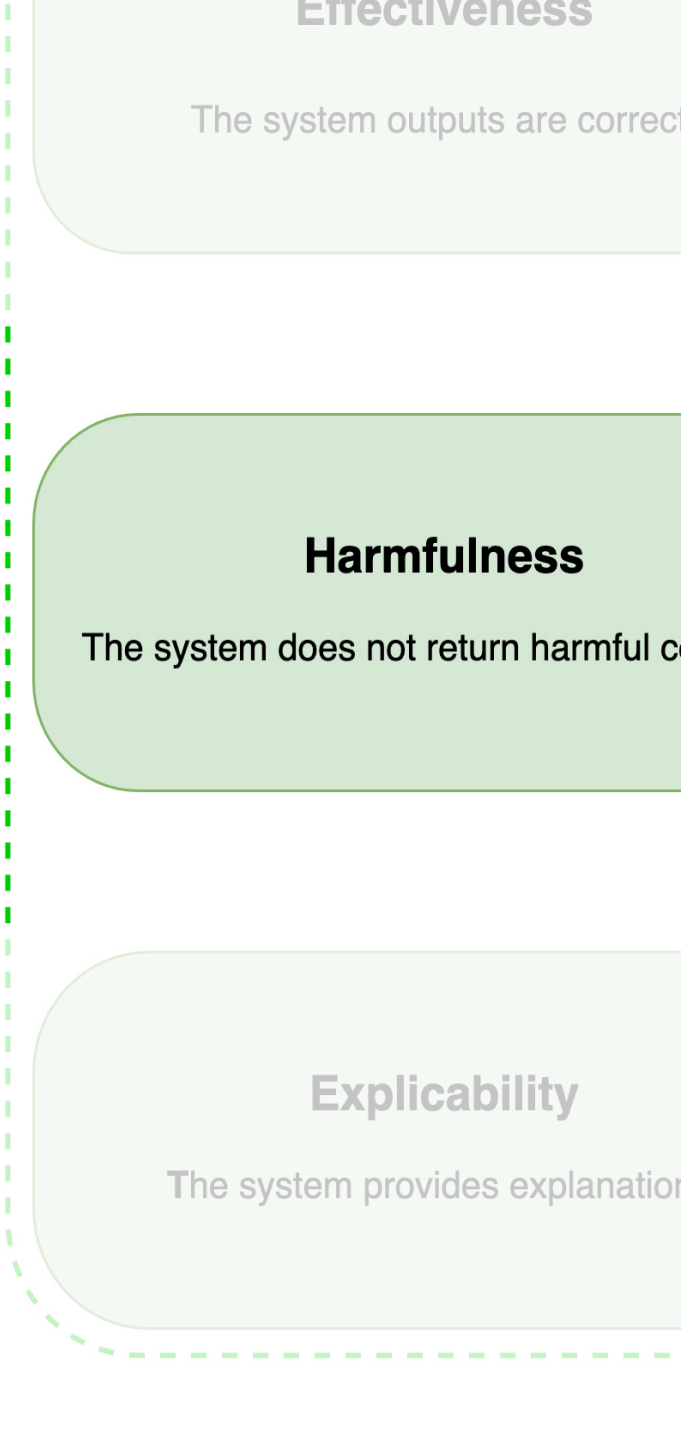
The system outputs are corr

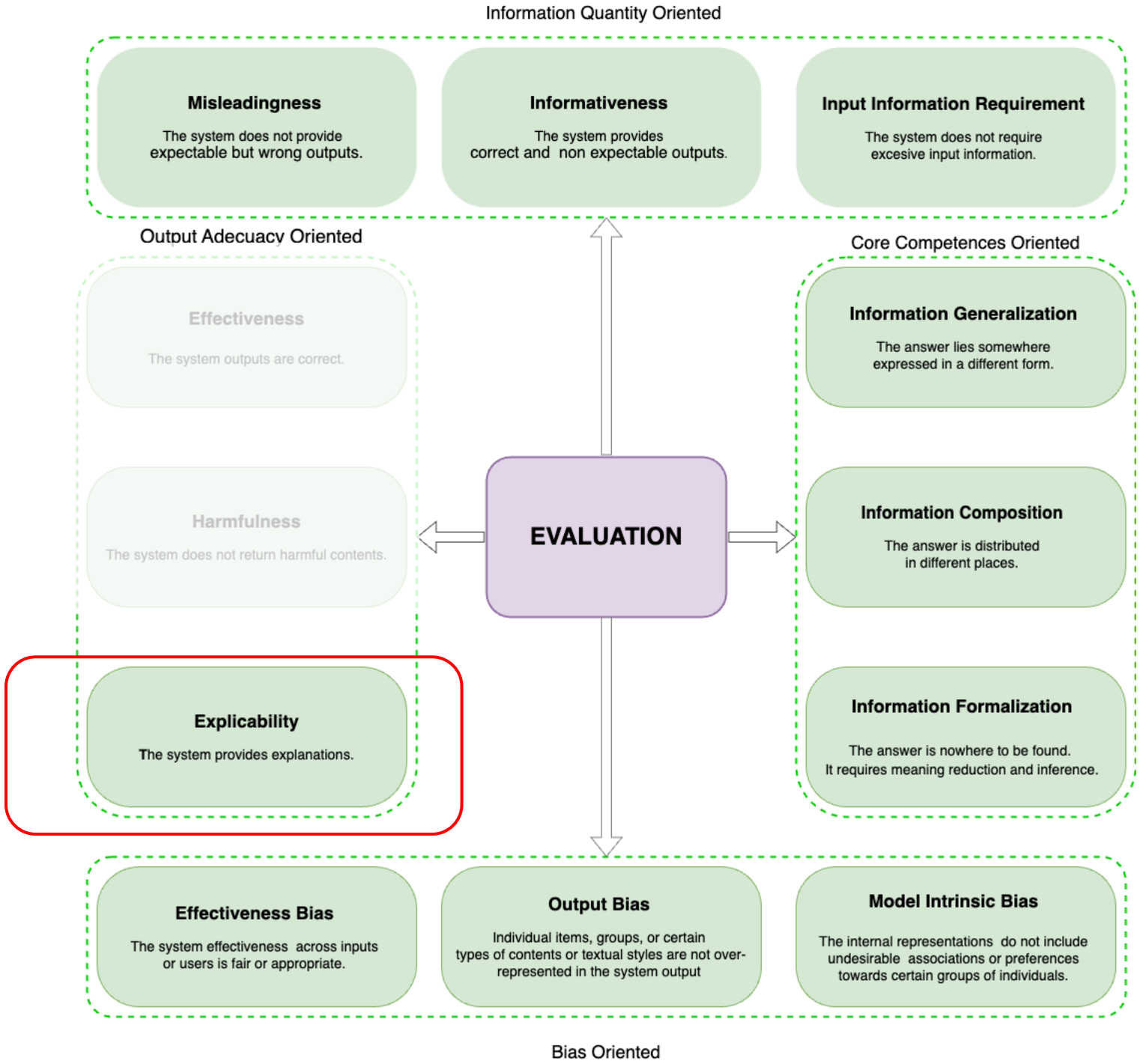
Harmfulness

The system does not return harmfu

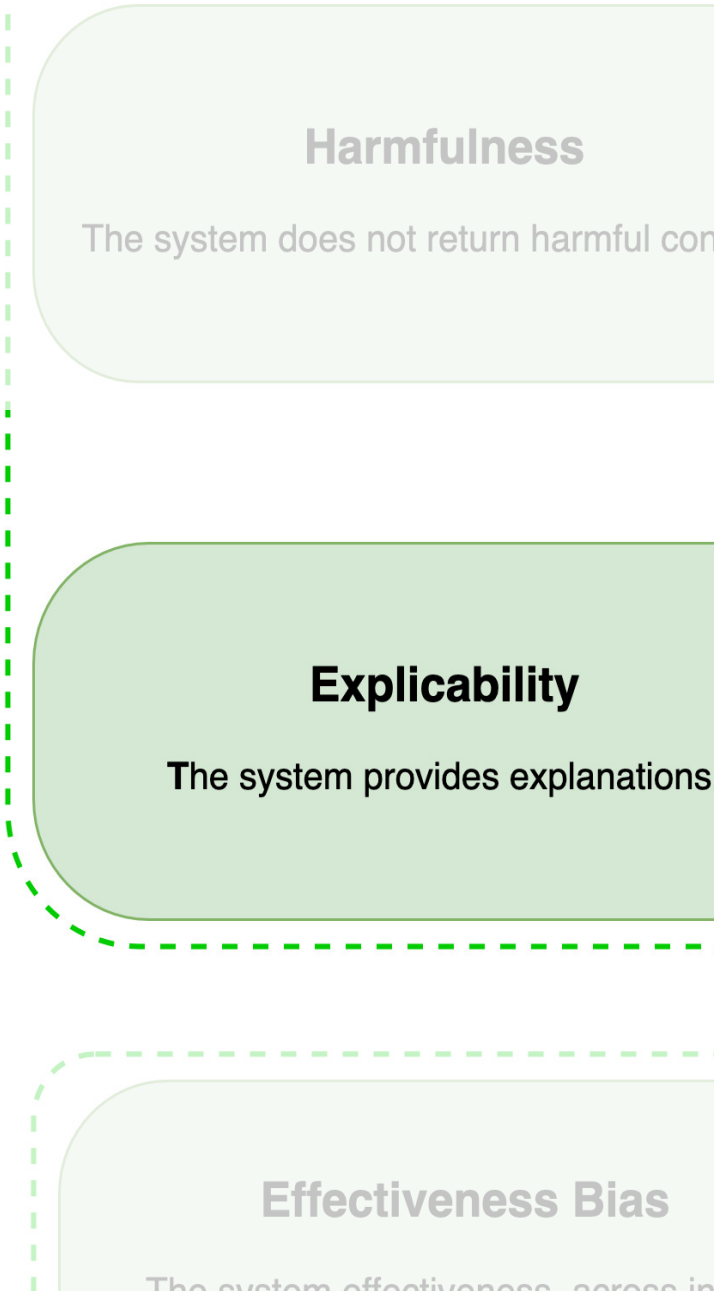

Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

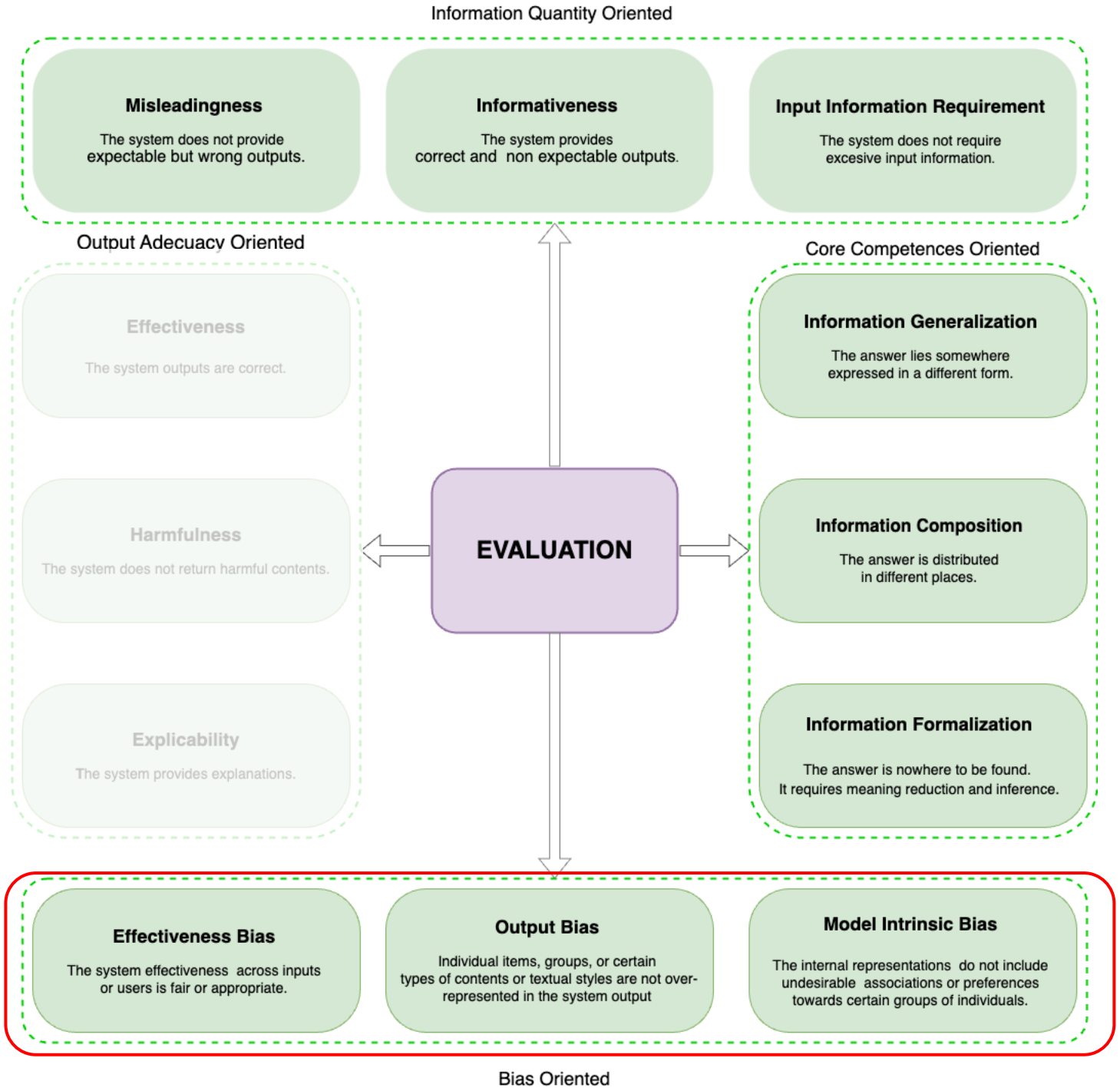
- Harmful decisions: Cost matrix in classification…
- Harmful contents:
  - Classification based:
    - General purpose: Sexism, hate speech, …
    - Trained on LLM safety data sets: (Zhang et al. 2024b)
  - Multiple choice tests: (Zhang et al. 2024a)
  - Prompting based: Provoking LLMs (Wullach et al. 2020, 2021), (Meyer et al. 2022b).




Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

- Component based (Labelling based task)
  - Tokens (Godin et al., 2018), alignments (Bahdanau et al., 2015), text sequences, (Mullenbach et al., 2018; Carton et al., 2018; Voskarides et al., 2015; Sydorova et al., 2019), training samples (Abujabal et al., 2017; Croce et al., 2019).
  - **Classification and sequence labelling metrics.** (e.g. F measure in (Carton et al., 2018))
- Formal explanations (Formal language)
  - Decision templates (Abujabal et al., 2017), knowledge graphs (Pezeshkpour et al., 2019), logic forms (Liang et al., 2016).
  - **Qualitative evaluation methods.**
- Textual explanations (Natural language generation)
  - Explanations in IR (Łajewska, et al 2024), mathematical problems (Ling et al., 2017), common sense questions (Rajani et al., 2019), etc.
  - **Text generation metrics** (BLUE, ROUGE, BertScore)




Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

Effectiveness Bias

The system effectiveness across inputs or users is fair or appropriate.

Output Bias

Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output

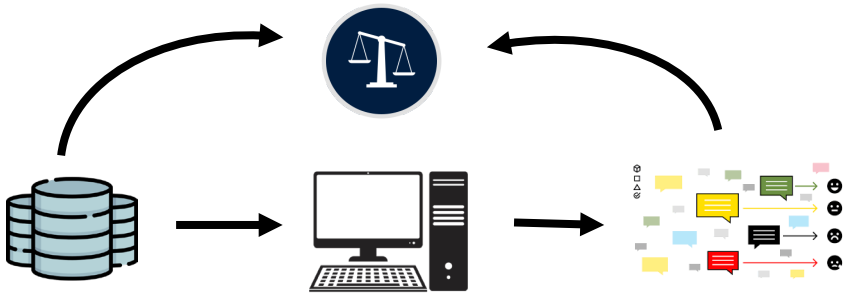
Model Intrinsic Bias

The internal representations do not include undesirable associations or preferences towards certain groups of individuals.

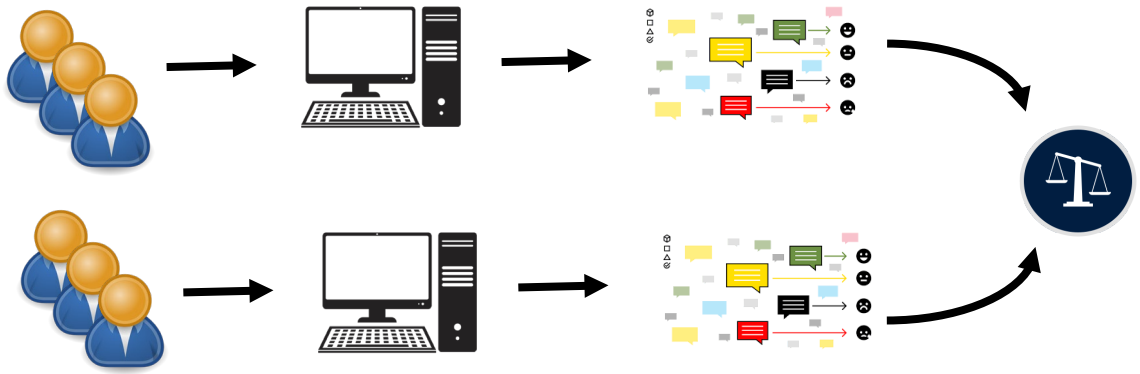
Bias: To what extent do the system favour or over-represent certain items, contents or users at an individual or group level.

What is being biased?

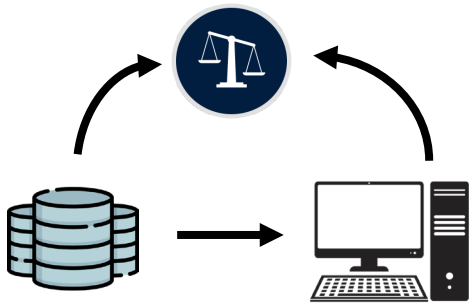
- **Output Bias:** Exposure distribution of items, categories, contents or linguistic features in the output.

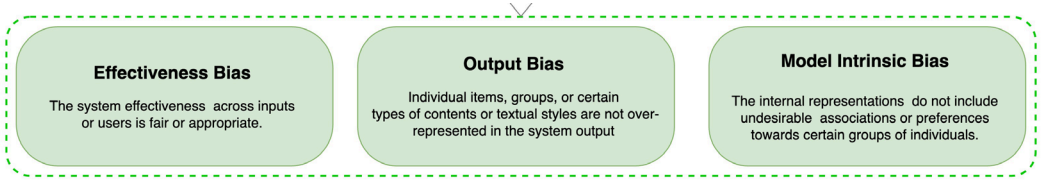


- **Effectiveness Bias:** Effectiveness across users.



- **Model Intrinsic Bias:**





Bias Criteria (What is the ideal distribution of outputs and effectiveness)

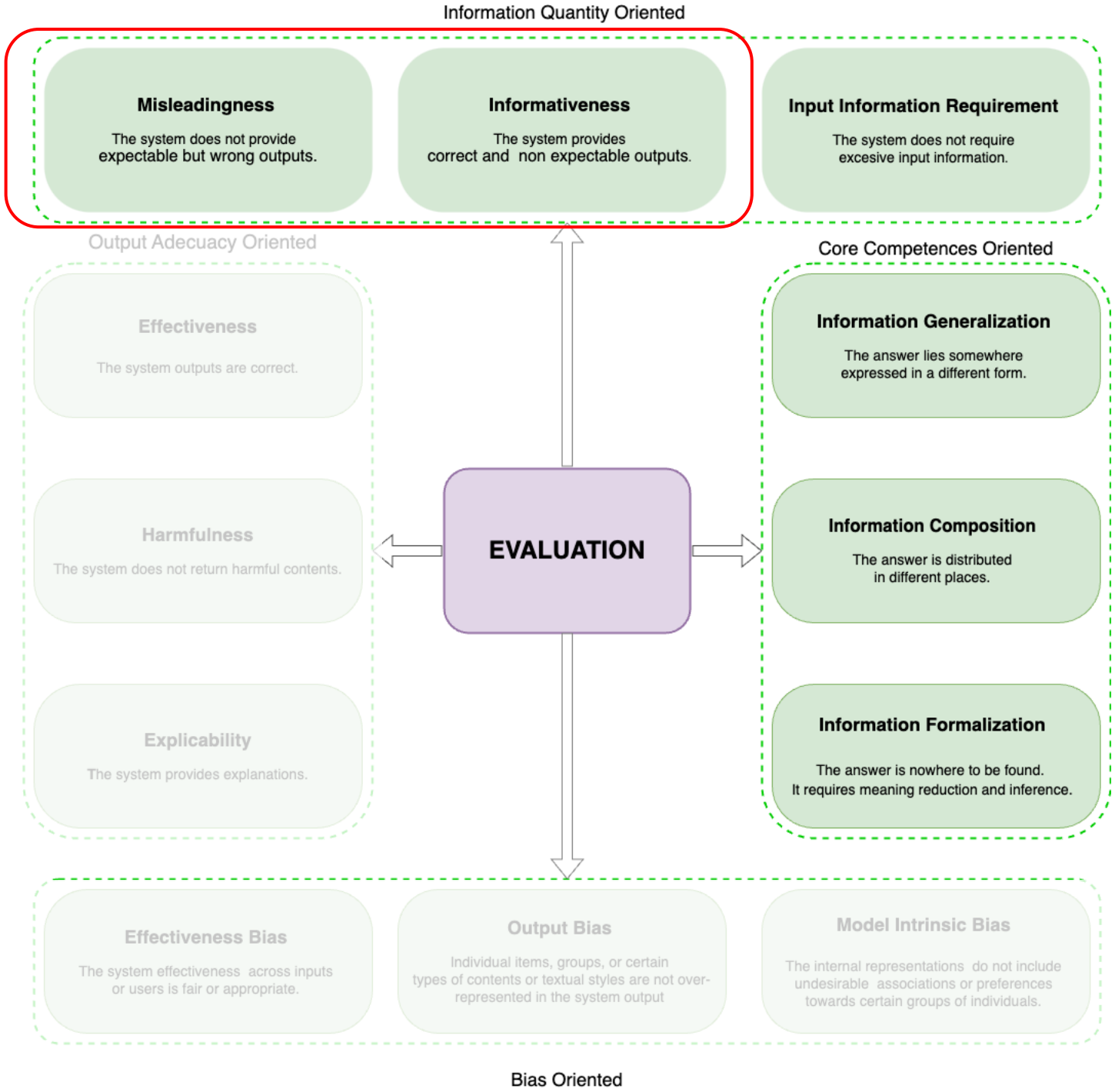
- Uniform: Equal benefit for all item/user individuals/groups.
- Proportional: Group size.
- Calibrated: According to individual/group characteristics (e.g. user group needs, recommended item group relevance)
- Consistency of individual treatment.
- Envy-free: Users do not prefer the output received by other users.
- Conterfactual: Certain sensitive attributes of an individual were changed while keeping other attributes constant.
- Rawlsian: The protection of the least advantaged members.
- Fairness through unawareness: No sensitive attributes are explicitly used in the decision-making process.

Model Intrinsic Bias

- Representation Similarity
  - Distance between representations associated with groups of individuals.
  - (Caliskan et al. 2017, May et al. 2019, Guo et al. 2021, Dolci et al. 2023)

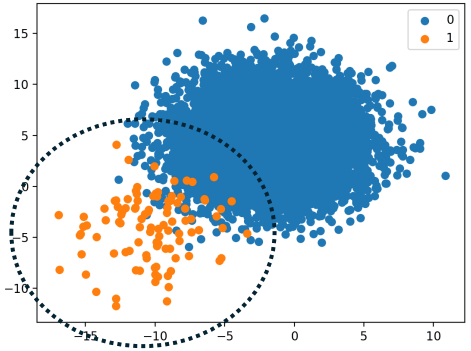
- Likelihood Measurement
  - Studying the probabilities assigned by the model to different words or sequences of words.
  - (Gallegos et al. 2024, Kurita et al. 2019, Nangia et al 2020, Nadeem et al. 2021, Kanekoet al 2021, Barikeri et al. 2021)

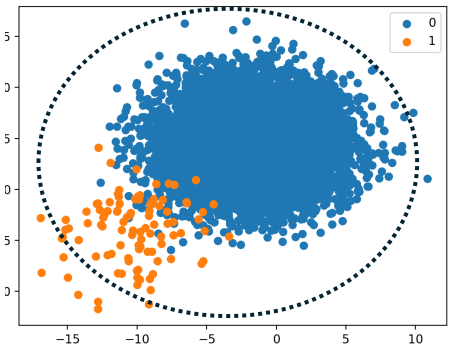
- Text Completion
  - comparing the model's response to two analogous prompts but associated with different population groups
  - Rajpurkar et al. 2016, Sicilia et al. 2023, Liang et al. 2023-holistic, Huang et al 2019, Sheng et al 2019, Smith et al. 2022, Nozza et al. 2021, Dhamalaet al 2021)


Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

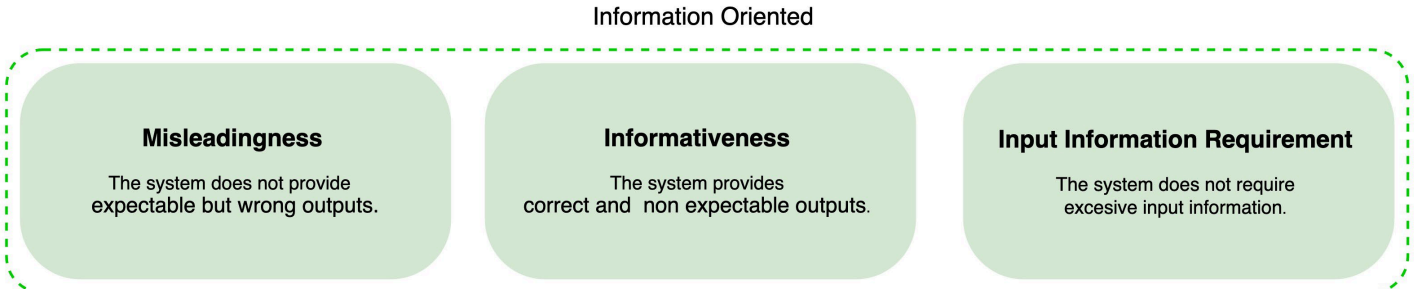
Informativeness and missleadingness

Organisational tasks: Classification and clustering



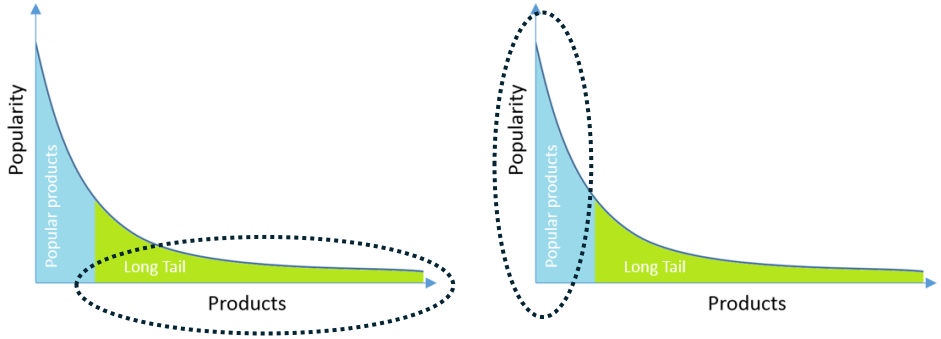


- **Informativeness**: Returning correctly minority labels.
- **Missleadingness**: Returning wrongly majority labels.
- Metrics:
  - Measuring effectiveness at class level (F measure, MAAC)
  - Information theory based metrics (ICM (Amigo and Delgado, 2022)), Entropy (Steinbach et al., 2000; Ghosh, 2003)), etc

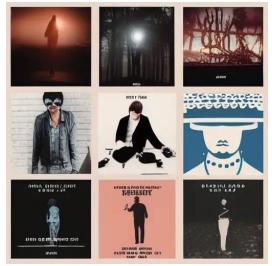


Informativeness and missleadingness

Information Retrieval and Recommendation.



- **Informativeness**: System ability to return unexpected but useful items.
- **Missleadingness**: The system returns expectable but unuseful items.
- Metrics
  - Recall based metrics (e.g. NDCG, MAP)
  - Diversification: Intent aware metrics (Agrawal et al., 2009).
  - Serendipity in recommendation: Metrics based on dissimilarity to a primitive system (Murakami et al., 2008; Ge et al., 2010)., or to the user's history (Zuva and Zuva, 2017)



Information Oriented

Misleadingness

The system does not provide expectable but wrong outputs.

Informativeness

The system provides correct and non expectable outputs.

Input Information Requirement

The system does not require excesive input information.

Misleadingness

Text Generation (Hallucination)

- Hallucination definition:
  - "*The generated content that is nonsensical or unfaithful to the provided source content* ." (Ji et al., 2023)
  - *"Fluent but unsupported text."* (Filippova, 2020)
  - *"Generating expectable but wrong responses."*
- Metrics: Comparing generated texts with sources.
  - Textual proximity (Dhingra et al., 2019, Shuster et al., 2021)
  - Information extraction tools (Goodrich et al., 2019)
  - Question Answering applied to the output vs. the sources (Durmus et al., 2020, Honovich et al., 2021).
  - Textual inference: (Dziri et al., 2022 , Huang et al., 2021; Kryscinski et al., 2020).

Informativeness

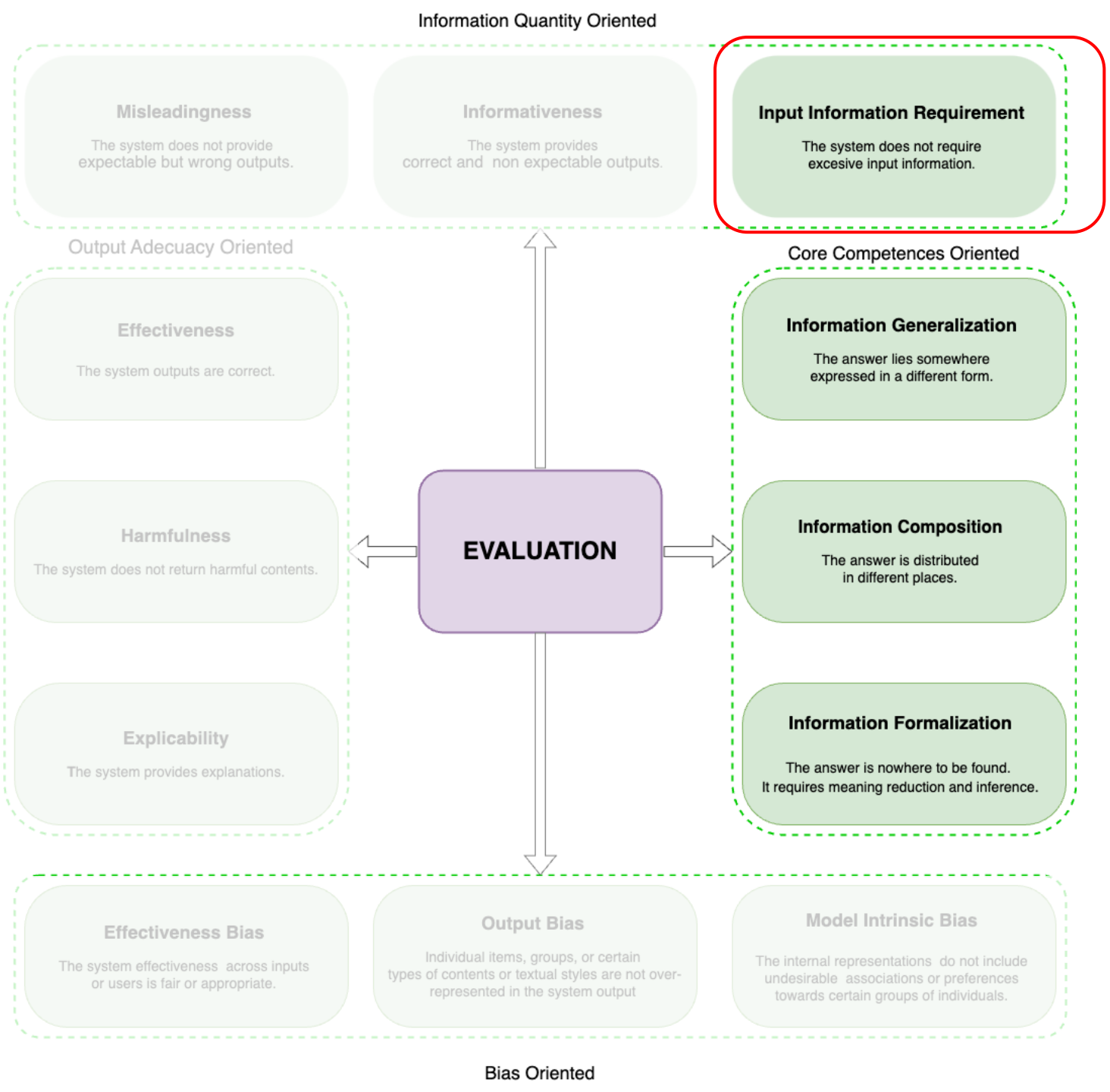
Text Generation

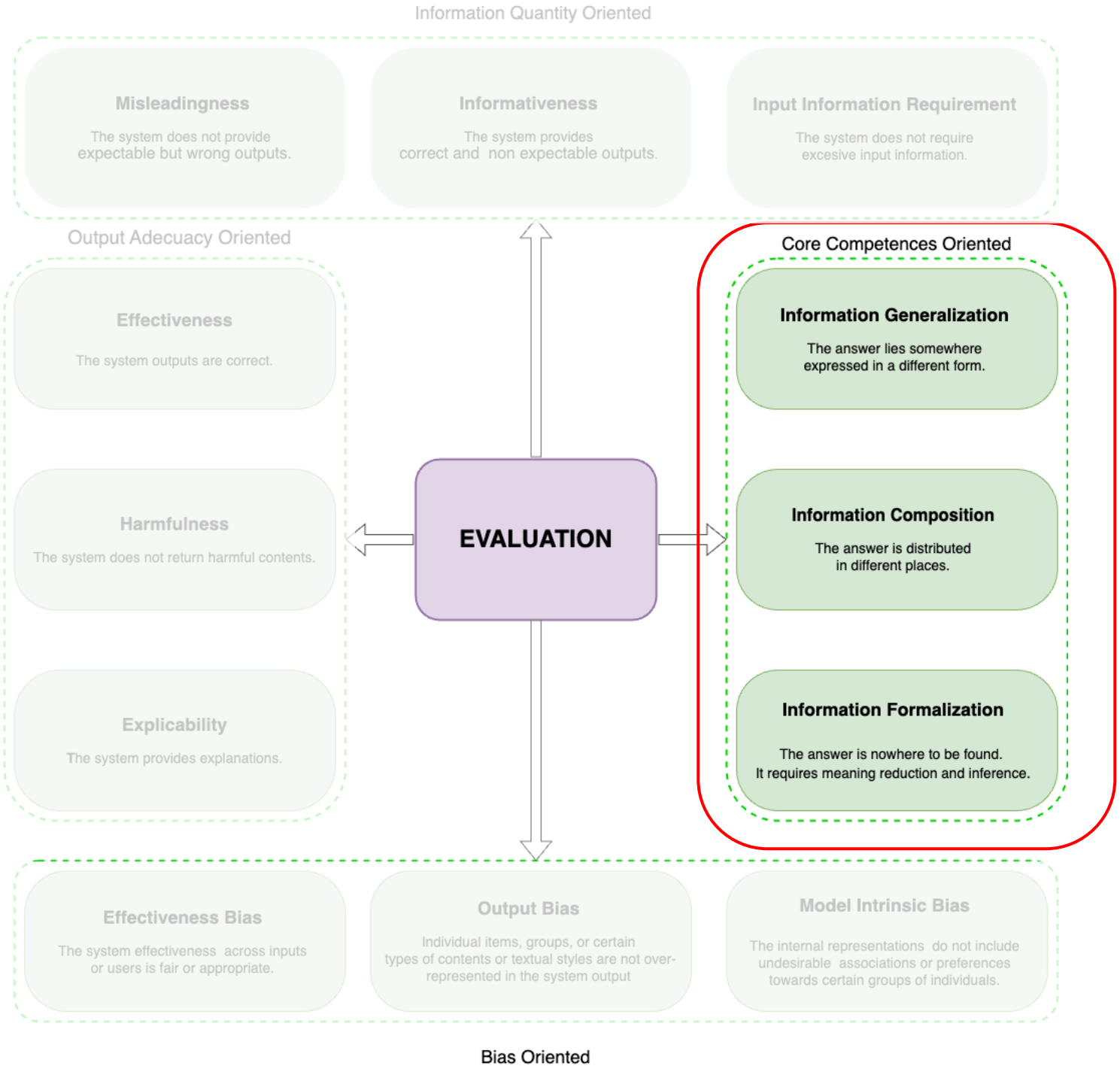
- *"Returning unexpected valuable response to an expected input"*
- All definitions of **creativity** converge on unexpectedness and effectiveness or usefulness.
- Unexpectedness depends on the reference probability distribution:
- Metrics and benchmarks:
  - Generating non related words: (Chen and Ding, 2023)
  - Human assessors (Marco et al., 2022, Summers-Stay et al., 2023).
  - Extrinsic evaluation (e.g. creative writing (Chakrabarty et al., 2023))
- *Should metrics combine correctness/adecuacy with dissimilarity with respect to the LM on which the system is based?*

Misleadingness vs Informativeness

Text Generation

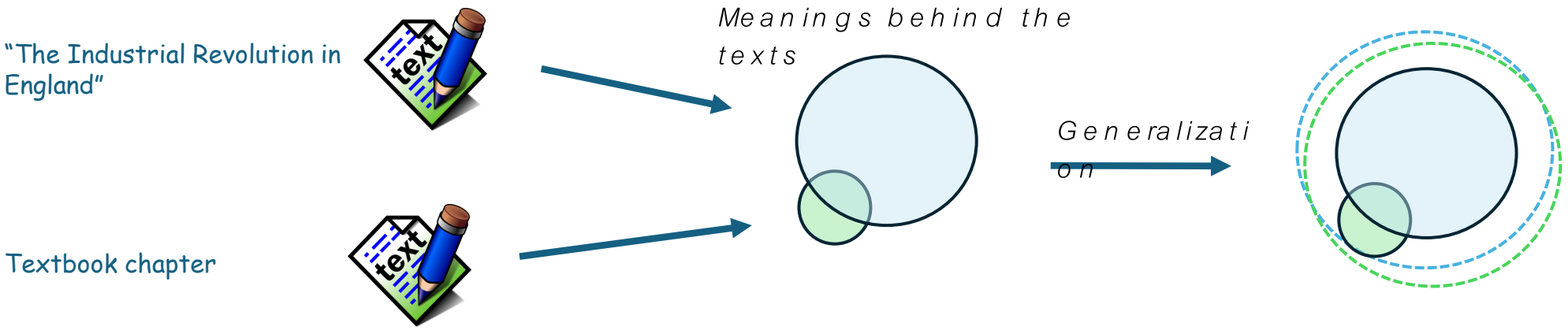
- Similarity to sources prevent hallucination.
- Dissimilarity to sources is an indicator of creativity.

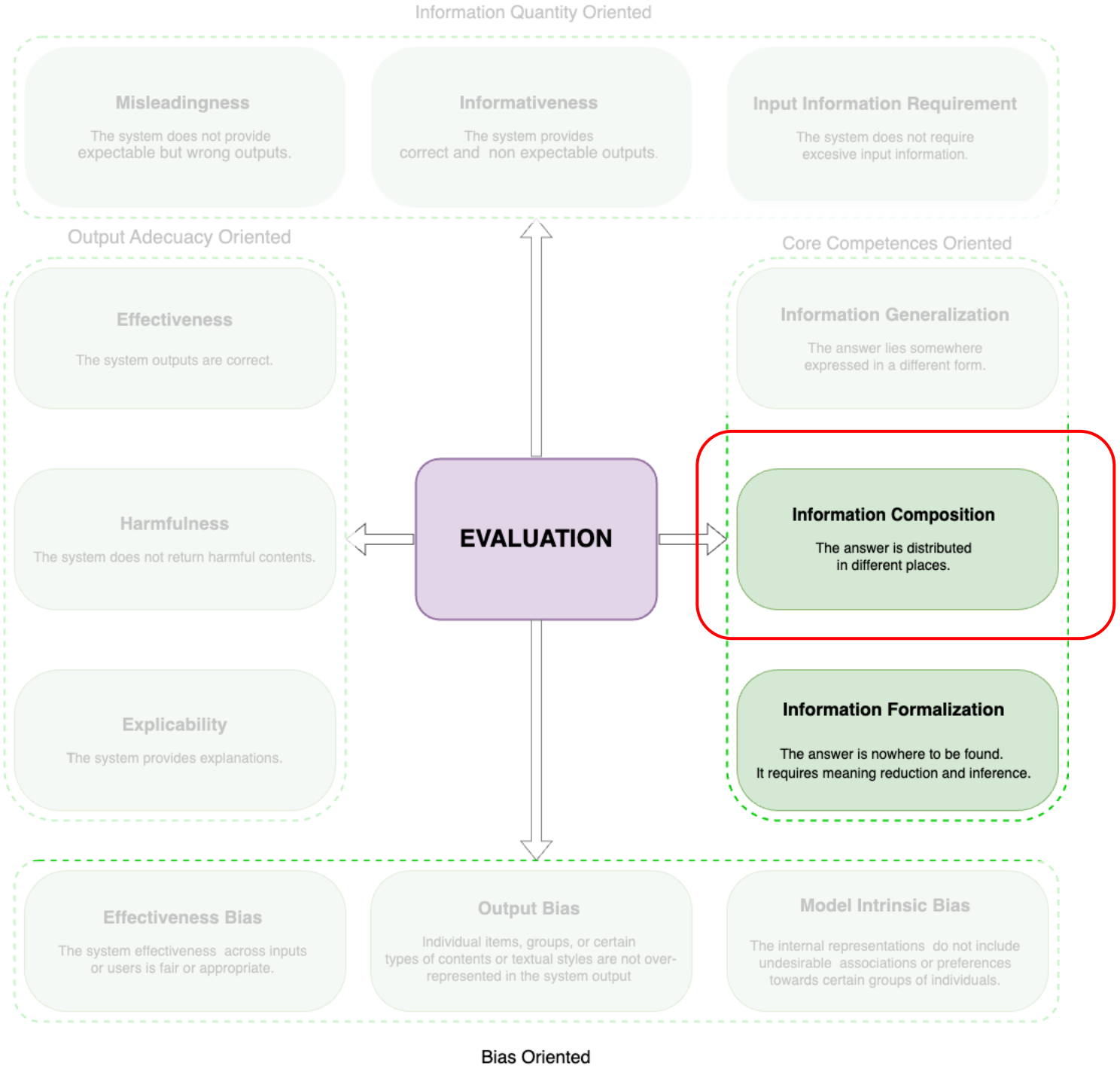

Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Core Competences Oriented
Effectiveness
The system outputs are correct.
Information Generalization
The answer lies somewhere expressed in a different form.
Harmfulness
The system does not return harmful contents.
EVALUATION
Information Composition
The answer is distributed in different places.
Explicability
The system provides explanations.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented


Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

Meaning Generalization

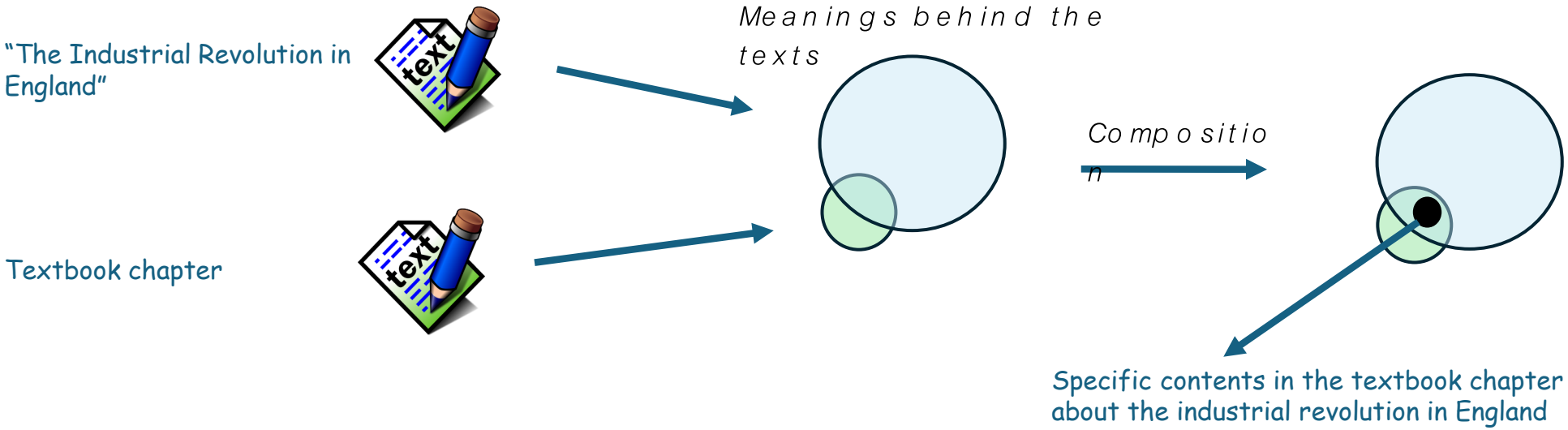
- The system is capable of generalizing meanings to connect similar information pieces.
- Benchmarks
  - Classification and Information retrieval: Traditional metrics
  - Text generation: Question typologies in Q&A. Traditional Q&A metrics. (TriviQA, Joshi et al., 2017).




Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

Meaning Composition

The system is capable of produce information by intersecting meanings.



Meaning Composition

Related concepts:

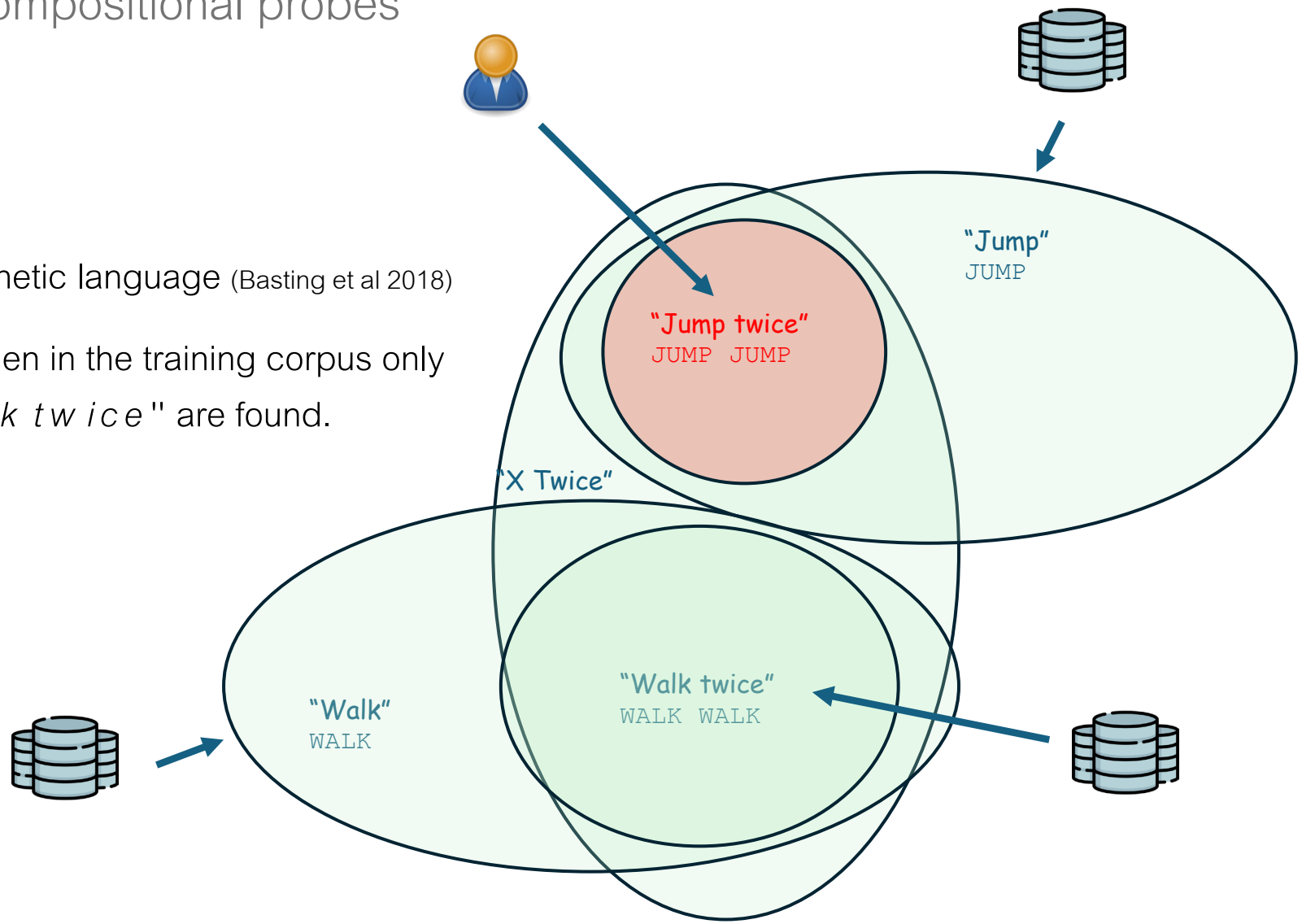
- **Compositional generalization:**
  - *"The ability to extend learning through the combination of individual elements into a more complex structure"* (Fodor and Lepore, 2002).
  - *"The ability to generalize to novel combinations of elements observed during training"* (Shaw et al., 2021).
  - *"Text processing when the corresponding formal expression is not explicitly present in the training data"* (Gu et al., 2020).
- **Linguistic compositionality:**
  - Systematicity, productivity, substitution, localism and overgeneralization (Hupkes et al., 2021).
  - Systematicity, substitutivity and global compositionality (Dankers et al 2022)
  - Primitive substitution, alternation of primitive structures, combination of phrasal constituents (An et al., 2023)
  - Negation, antonymy, entity substitution, mutual exclusion, or impossible conditions (Rajpurkar et al., 2018)

Meaning Composition: Evaluation benchmarks

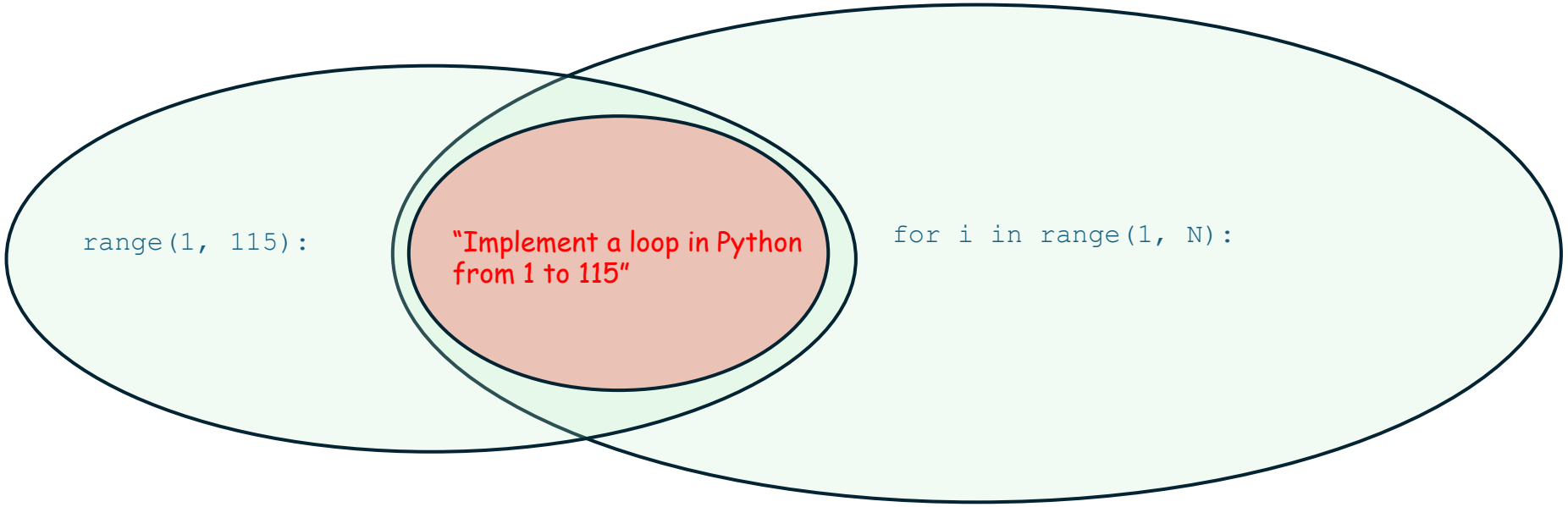
- Meaning compositional tasks
  - Simole code generation (Chen et al., 2021; Austin et al., 2021a)
    - Formal language generation
    - Code evaluation metrics
  - Text to SQL (Saeed et al., 2023).
  - Long-Form Q&A (Qin et al., 2023)
    - Text generation metrics.
  - Reading and comprehension (Rajpurkar et al., 2016; Choi et al., 2018; Trischler et al., 2017).
    - Answers requiring information synthesis from multiple passages
    - Discrimination task: multi-choice, sequence labelling.
    - F-measure, accuracy, sequence labelling metrics.
- Compositionality probes
  - Compositionality oriented Natural Language Inference (Goodwin et al. 2020)
  - Instruction sequences: SCAN (Lake and Baroni, 2017), NACS (Bastings et al., 2018). Accuracy
  - Logic forms: CFQ (Keysers et al., 2020) COGS (Kim and Linzen, 2020). Accuracy

Meaning Composition: Compositional probes

- **Sca n** benchmark of synthetic language (Basting et al 2018)
- Interpret "*ju mp t w i c e* " when in the training corpus only "*w a l k*", "*ju mp* ", and "*w a l k t w i c e* " are found.

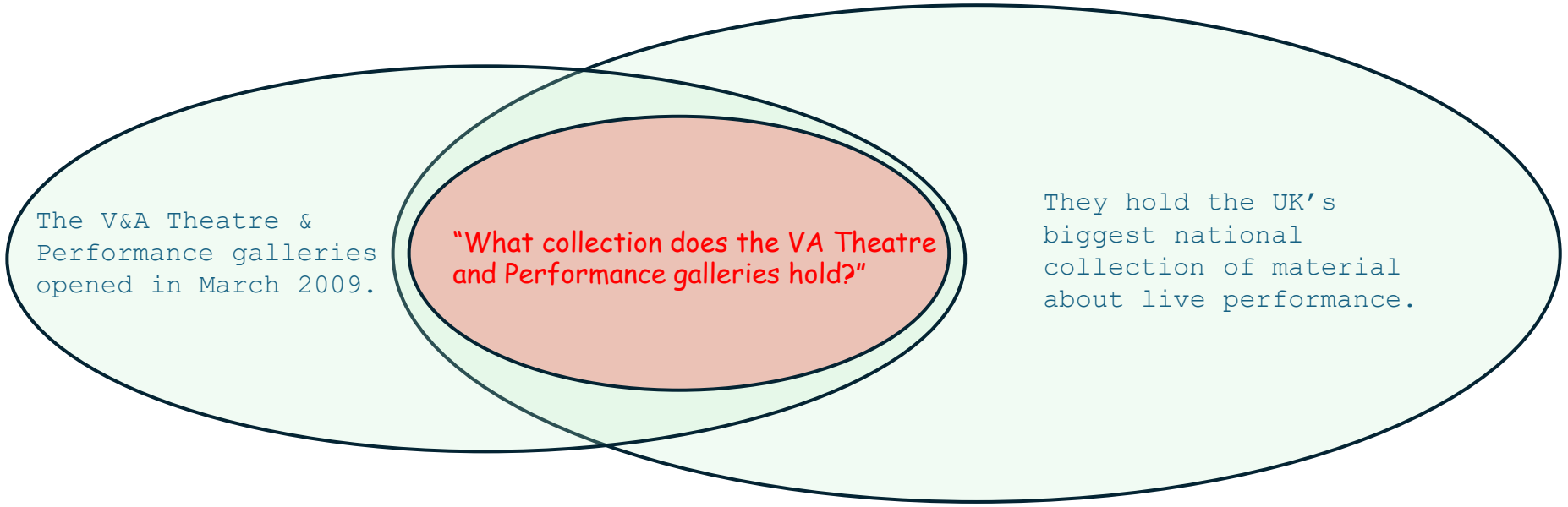


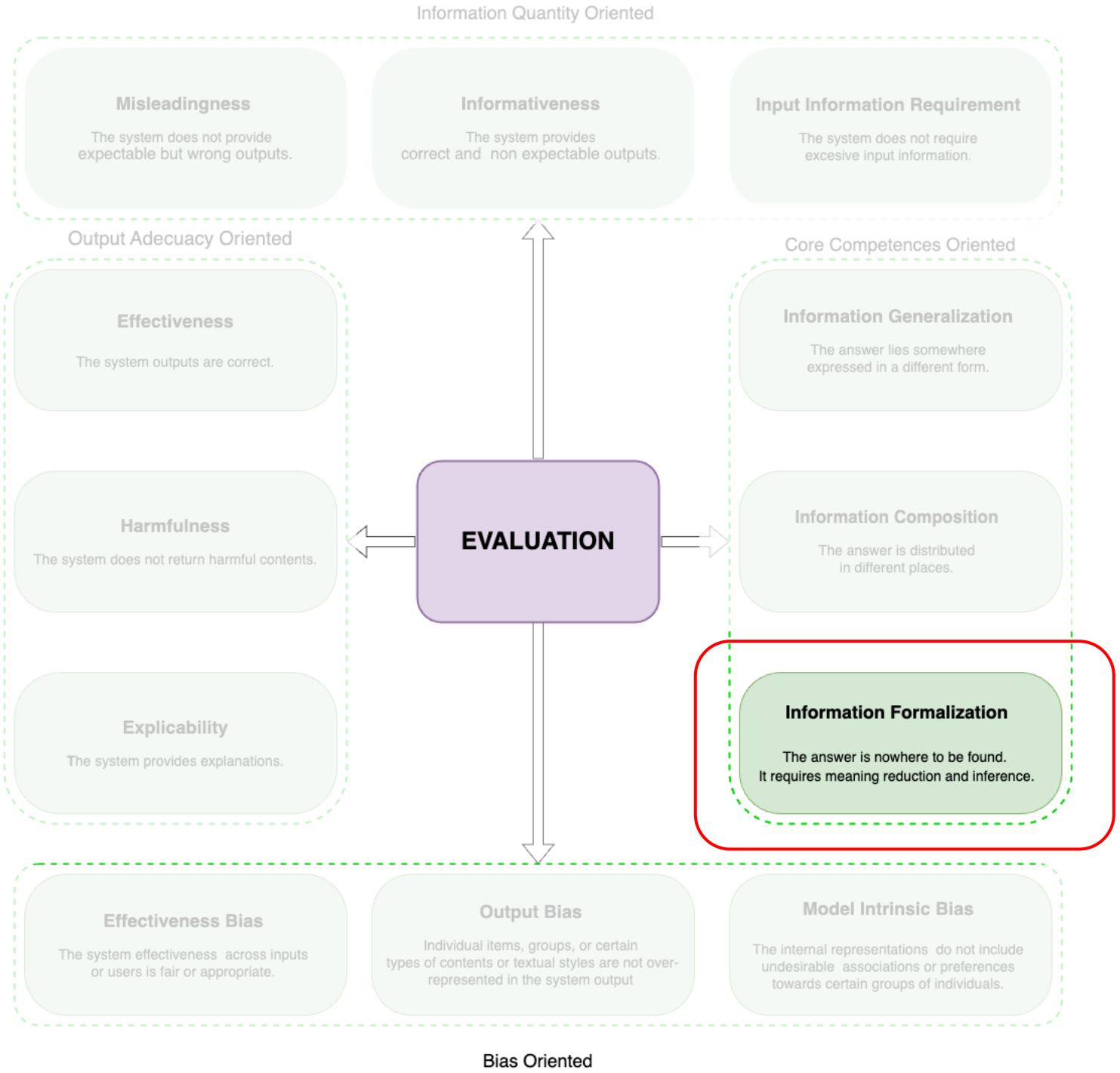
Meaning Composition: Code generation



Meaning Composition: Reading and comprehension

SQuAD: (Rajpurkar et al., 2016)



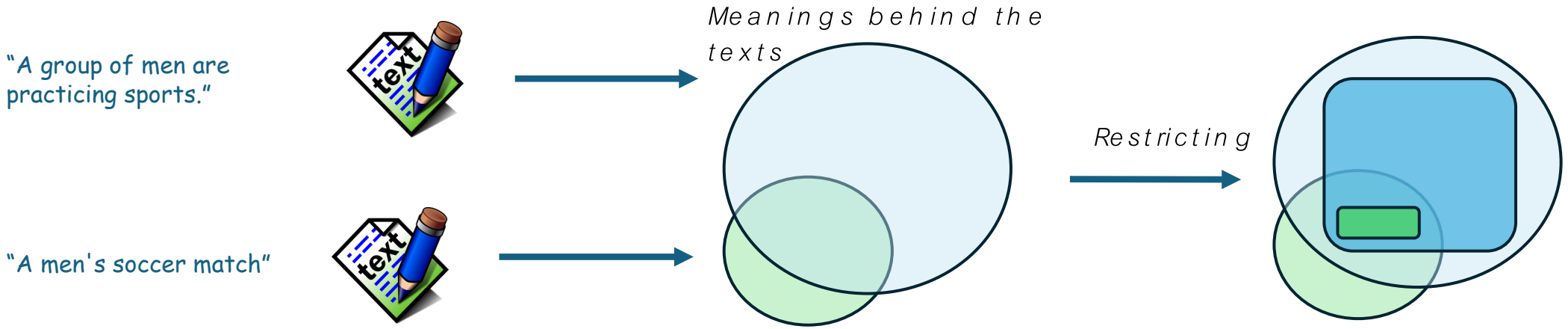

Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found.
It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

Meaning Formalization

The system is capable of produce information by restricting meanings (world modelling).

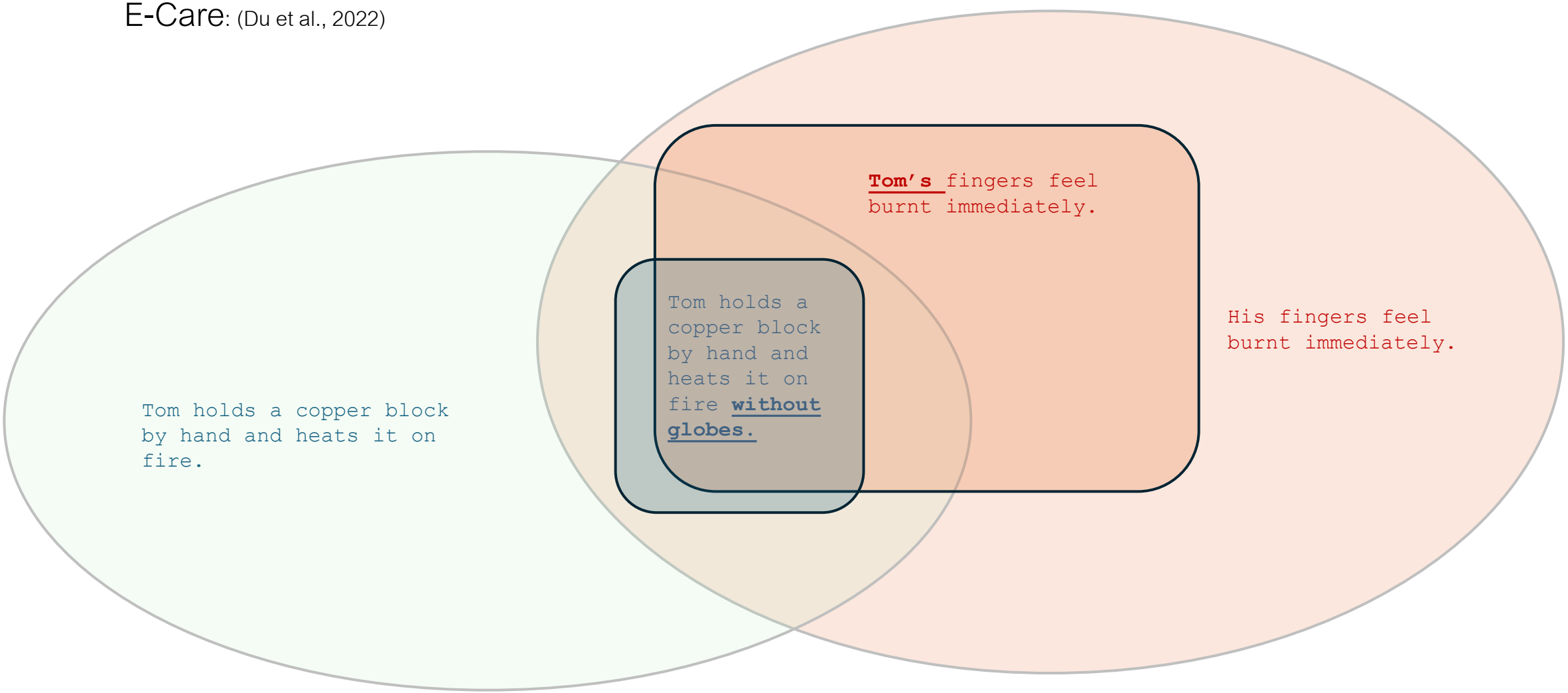
Problems that involve textual implication, planning, or mathematical reasoning.

Assumptions:
- It is not a computer soccer game.
- The players are not planning the game.

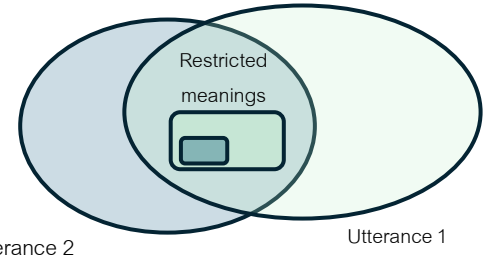


Meaning formalization: An example

E-Care: (Du et al., 2022)

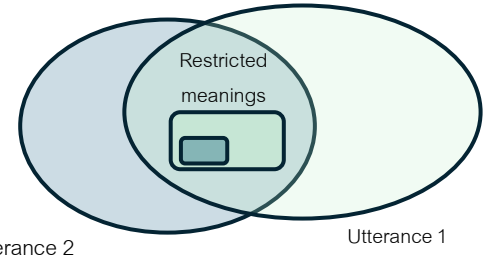


Meaning formalization: Benchmarks

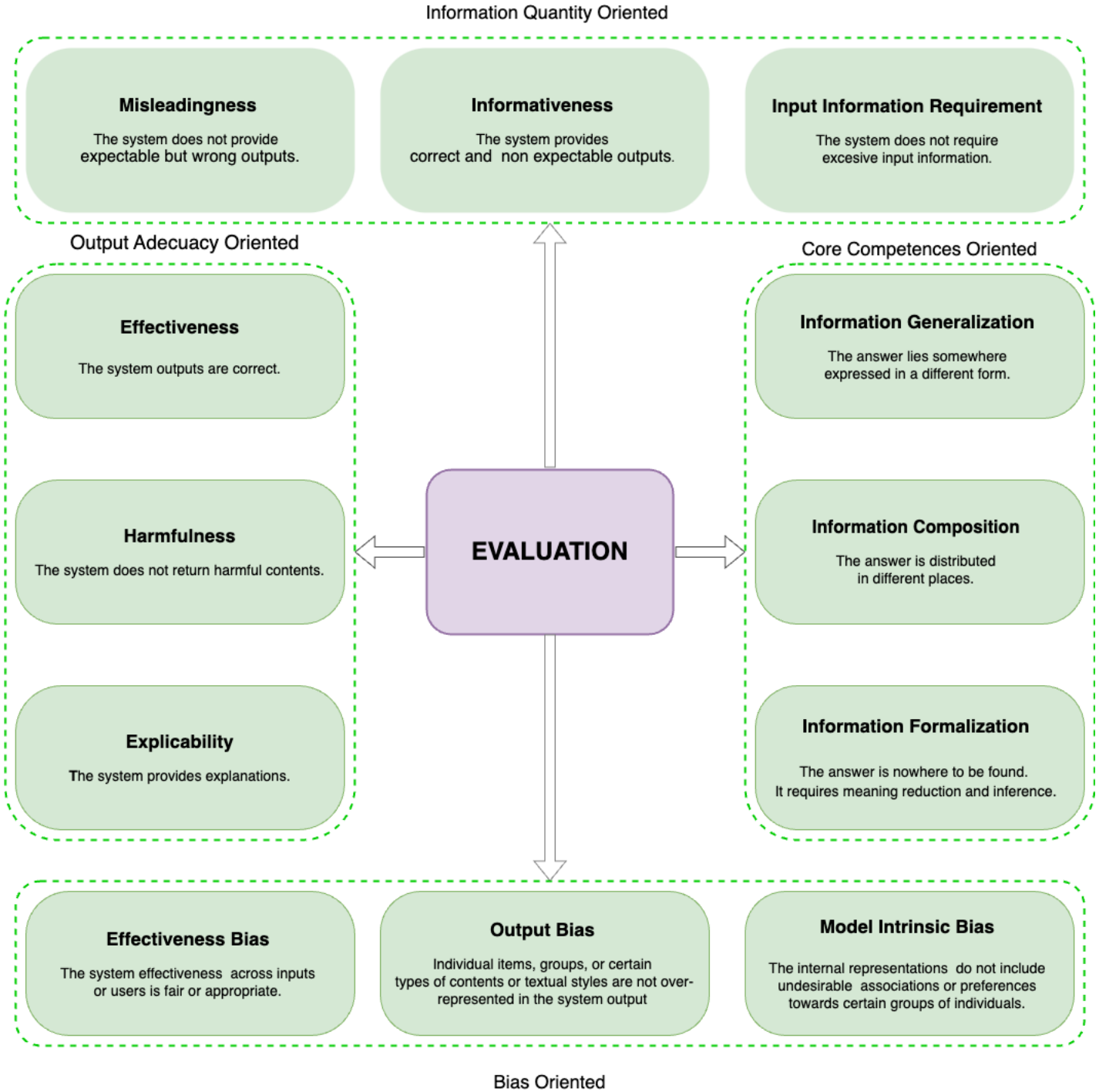


- **Identifying logic relationships:** SNLI (Bowman et al., 2015), Multi-NLI (Williams et al., 2018), ConjNLI (Saha et al., 2020 HELP (Yanaka et al., 2019). CausalBank (Li et al., 2021b), Abductive reasoning (Bhagavatula et al., 2020). FOLIO (Han et al., 2022). **(Accuracy based metrics)**
- **Multiple choice:** **(Accuracy based metrics)**
  - Event plausibility (Wang et al., 2018b), physical actions PIQA (Bisk et al., 2020), temporal reasoning (Zhou et al., 2021), causality (Du et al., 2022), social situations; Social IQA (Sap et al., 2019) Common sense; Swag (Zellers et al 2018) Reading Comprehension; COSMOSQA (Huang et al. 2019) WinoGrande (Sakaguchi et al. 2019) Entity deduction via question games (Zhang 2024)
  - Ensuring the need for reasoning: Knowledge graphs (Talmor et al., 2019), combining facts (Mihaylov et al., 2018, Khot et al., 2020), temporal and arithmetic reasoning in a dialog (Qin et al., 2021), pruning data sets with low system performance (Suzgun et al. 2023)
- **Mathematic outputs:** MATH (Srivatava et al, 2024) TabMWP (Lu et al., 2022b) GSM8k (Cobbe 2021) RGSM (Chen 2024) Math23K (Wang et al., 2017) and HMWP (Qin et al., 2020) **(Exact matching metrics)** DRAW1K dataset (Upadhyay and Chang, 2017) **(Derivation accuracy)**
- **Language generation:** CommonS Sense: CommonGen (Lin et al., 2020b) **(Text evaluation metrics)** Logic problem (Ontañón et al., 2022). **(Exact matching metric),** Logic reasoning justification (Du et al., 2022). **(Lexical overlap metric based on causal strength),** CoT on ontologies. (Saparov et al, 2024) **(Semantic parsing metrics)**

Some notes about core competence evaluation

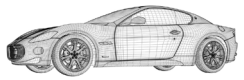


- Core competence levels are accumulative.
- Core competence depends on the training data.
- High core competences can be achieved by combining systems with lower core competences.
- High core competences (formalization and world modelling) allow to return unexpected, non biased, explainable, and non harmful responses.

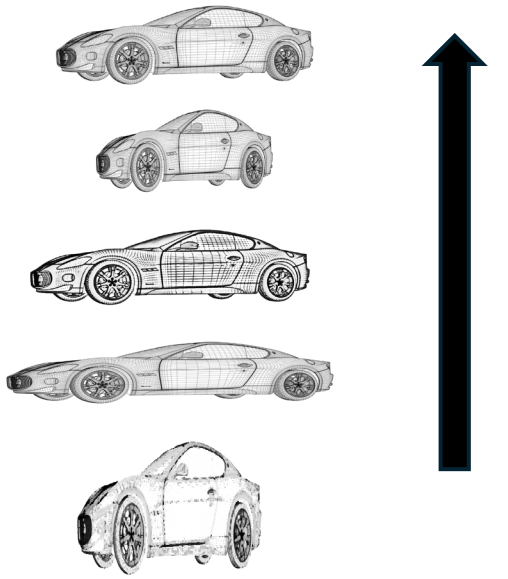

Information Quantity Oriented
Misleadingness
The system does not provide expectable but wrong outputs.
Informativeness
The system provides correct and non expectable outputs.
Input Information Requirement
The system does not require excesive input information.
Output Adecuacy Oriented
Effectiveness
The system outputs are correct.
Harmfulness
The system does not return harmful contents.
Explicability
The system provides explanations.
EVALUATION
Core Competences Oriented
Information Generalization
The answer lies somewhere expressed in a different form.
Information Composition
The answer is distributed in different places.
Information Formalization
The answer is nowhere to be found. It requires meaning reduction and inference.
Effectiveness Bias
The system effectiveness across inputs or users is fair or appropriate.
Output Bias
Individual items, groups, or certain types of contents or textual styles are not over-represented in the system output
Model Intrinsic Bias
The internal representations do not include undesirable associations or preferences towards certain groups of individuals.
Bias Oriented

What can't we do?

Quality $\left(\quad \right) = \int_a^b \frac{x + x^3}{\alpha x^4 + \left[\frac{x + 2x^3}{3x - x^4} \right]^4} dx$

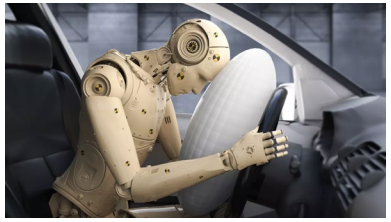


Car model ranking



What can we do?

Simulating users in quality tests



Competitions in evaluation campaigns

Multiple metrics and tests for multiple components and quality dimensions