

Testing Modelling Fitness of Normative Specification Languages for Large Language Models

NLL2FR Workshop, joint with JURIX 2025,
Torino, 9 December, 2025



Giovanni Sileno
g.sileno@uva.nl

Andrea Marino
a.marino@uva.nl

University of Amsterdam

The Babel tower of norm specification languages

- The number of languages for norm specification continues to increase in the literature...



The Babel tower of norm specification languages

- The number of languages for norm specification continues to increase in the literature...
- In the last 10 years only, we can find:
Defeasible Deontic Logic/SPINdle,
LegalRuleML, NPL/NPL(s), PROLEG, InstAL,
ODRL, Symboleo, FLINT/eFLINT, Logical
English, Catala, Blawx, Stipula, DCPL...



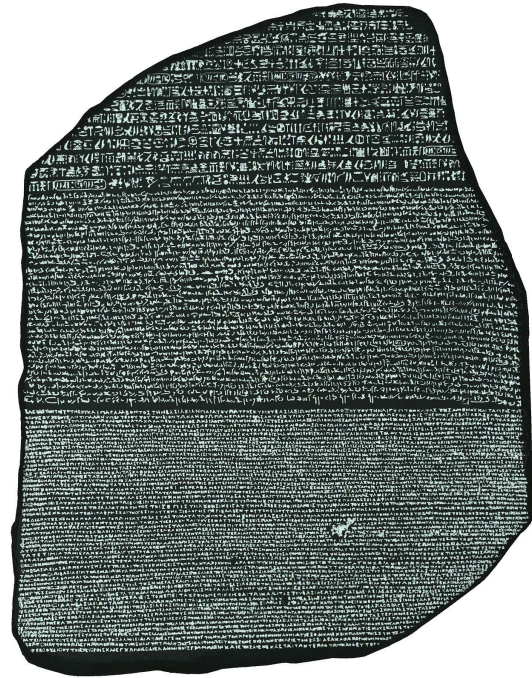
The Babel tower of norm specification languages

- The number of languages for norm specification continues to increase in the literature.
- In the last 10 years only, we can find: Defeasible Deontic Logic/SPINdle, LegalRuleML, NPL/NPL(s), PROLEG, InstAL, ODRL, Symboleo, FLINT/eFLINT, Logical English, Catala, Blawx, Stipula, DCPL...
- Furthermore, we have also technical-centred languages (access-control, usage-control, networking, cloud, ...), and works on alignment (safety, probabilistic shields, ...), which aim to regulate computational actors.



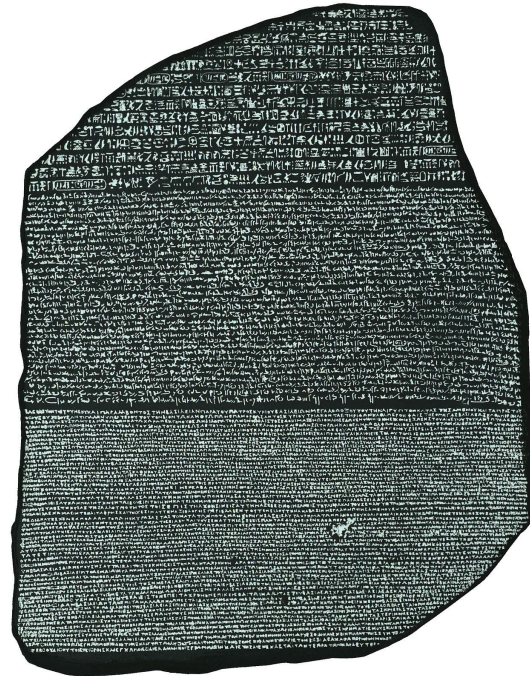
Comparison is difficult, because...

- Traditionally, comparison of formal languages is made in terms of their expressive power, yet this is a challenging task, also because they have been constructed from **different formal traditions** and for **different purposes**.



Comparison is difficult, because...

- Traditionally, comparison of formal languages is made in terms of their expressive power, yet this is a challenging task, also because they have been constructed from **different formal traditions** and for **different purposes**.
- (For a comparison framework including socio-technical dimensions, see Sofia Ali on Wednesday, Session III 14:00–15:30 main JURIX conference).

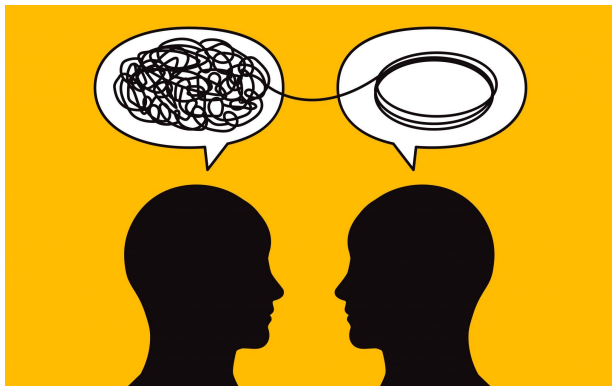


Applying LLM for the knowledge bottleneck?

- Efforts to use LLMs to support the translation of norms into computational processable notations are increasing in number...

Applying LLM for the knowledge bottleneck?

- Efforts to use LLMs to support the translation of norms into computational processable notations are increasing in number...
- *Which target language provides the most appropriate mapping for converting natural language into a formal notation?*



Measuring “Modelling fitness”?

- In biology, ecological fitness is the capacity of an individual or a population to maintain its functionality over time within an ecosystem.

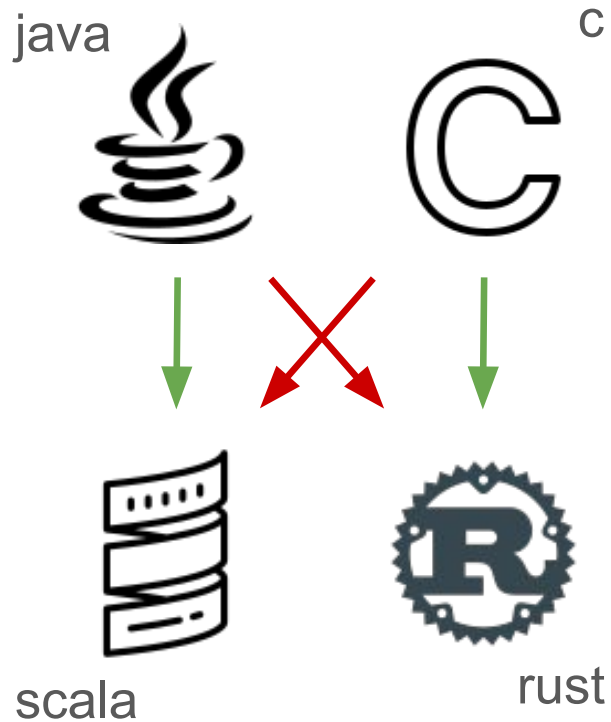
Measuring “Modelling fitness”?

- In biology, ecological fitness is the capacity of an individual or a population to maintain its functionality over time within an ecosystem.
- We similarly introduce **modelling fitness** to denote the *viability of a notation for a given modelling task*, assessed relatively to the modeller.

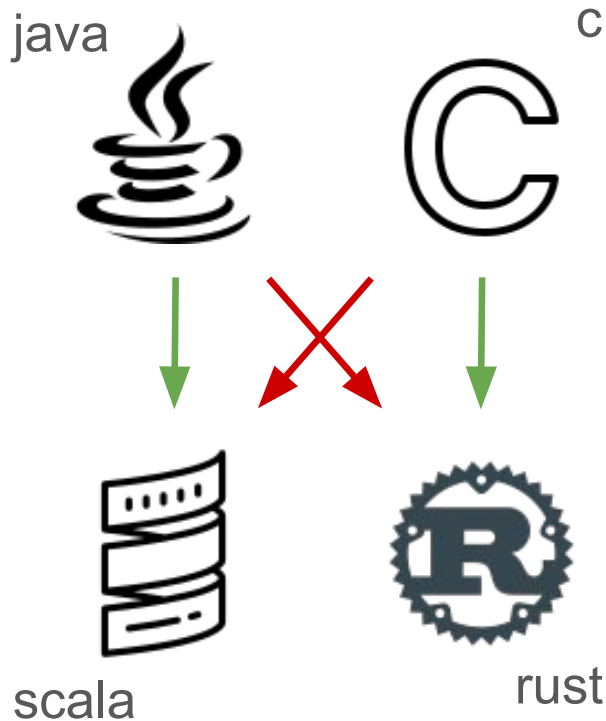
Measuring “Modelling fitness”?

- In biology, ecological fitness is the capacity of an individual or a population to maintain its functionality over time within an ecosystem.
- We similarly introduce **modelling fitness** to denote the *viability of a notation for a given modelling task*, assessed relatively to the modeller.
- For humans, it has a high variability!

For instance, for a programmer...



For instance, for a programmer...



Note that this is not a matter of expressivity!
(They are all general purpose programming languages!)

Main components of modelling fitness

1. capacity for a correct interpretation of the scenario at stake

depending on natural language proficiency and domain knowledge of the user...

2. proficiency with the modelling notation

depending on the accessibility and usability of the notation for the user's conceptual resources...

Main components of modelling fitness

1. capacity for a correct interpretation of the scenario at stake

depending on natural language proficiency and domain knowledge of the user...

2. proficiency with the modelling notation

depending on the accessibility and usability of the notation for the user's conceptual resources...

and for LLMs?

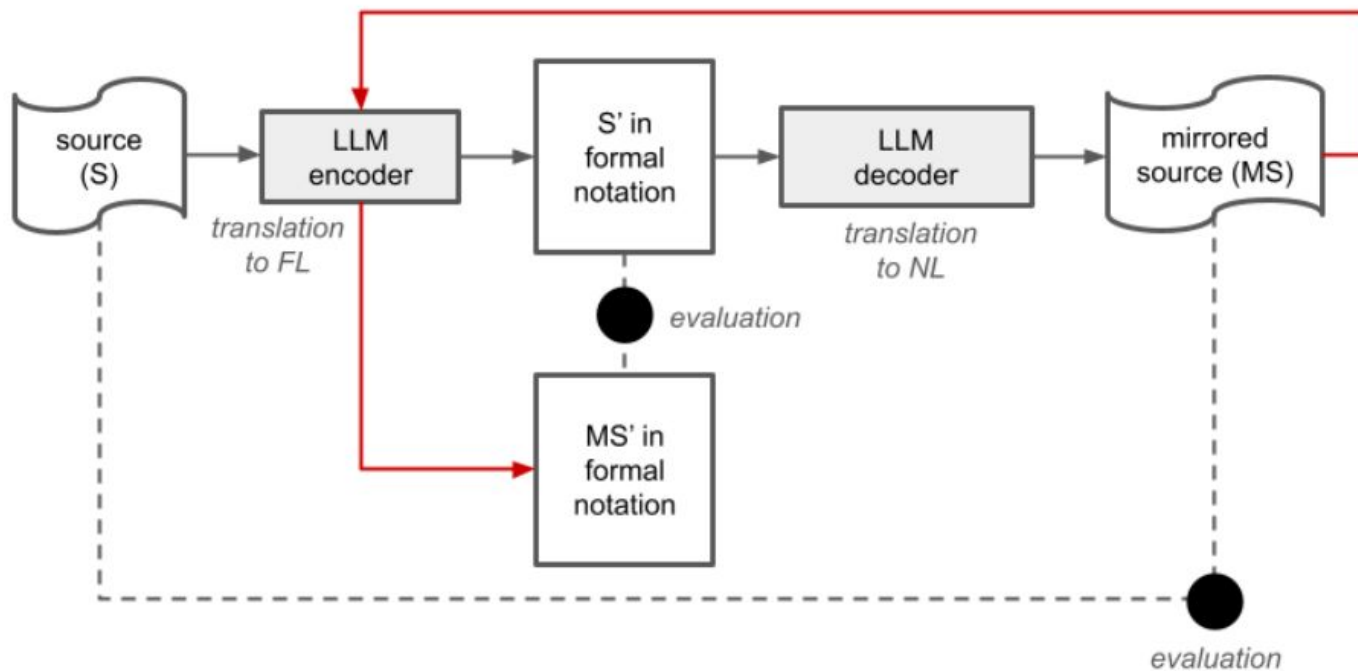
Evaluation method (1)

- Most LLM-supported translations are evaluated against reasoning tasks. Yet, some notations (eg. ODRL, a W3C standard) provide only an informational model.



we introduce the ***mirroring pipeline***, inspired by the autoencoder architecture

Mirroring pipeline



Evaluation method (2)

- Most LLM-supported translations are evaluated against reasoning tasks. Yet, some notations (eg. ODRL, a W3C standard) provide only an informational model.
- Some of these languages are niche. We provide with the prompt grammar (BNF) or informational model (eg. JSON-schema), or few examples of use, or both.



We perform ablation tests to see individual contributions.

Evaluation method (3)

- Most LLM-supported translations are evaluated against reasoning tasks. Yet, some notations (eg. ODRL, a W3C standard) provide only an informational model.
- Some of these languages are niche. We provide with the prompt grammar (BNF) or informational model (eg. JSON-schema), or few examples of use, or both.
- We aim to test the various languages across a wide spectrum of normative specification scenarios (covering plausibility and specificity, eg. normative primitives or simple compositions).



In this paper we tested 48 small scenarios taken from the literature, education material and technical documents.

Results DCPL vs ODRL

deep equality of
the two FN outputs

structural equality

Model	Language	Semantic score	(a)	(b)	(c)	(d)	$a \cdot b \cdot c$	$a \cdot b \cdot c \cdot d$
gpt5-mini	DCPL	0.82 ± 0.12	0.92 ± 0.25	0.76 ± 0.26	0.86 ± 0.27	0.83 ± 0.27	0.68 ± 0.27	0.60 ± 0.28
	ODRL	0.64 ± 0.20	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.49 ± 0.35	1.0 ± 0.0	0.49 ± 0.35
gpt4.1	DCPL	0.83 ± 0.14	0.74 ± 0.44	0.62 ± 0.40	0.67 ± 0.42	0.67 ± 0.42	0.56 ± 0.38	0.50 ± 0.36
	ODRL	0.75 ± 0.17	0.97 ± 0.16	0.94 ± 0.18	0.97 ± 0.16	0.51 ± 0.29	0.94 ± 0.18	0.49 ± 0.29
gpt4.1-mini	DCPL	0.78 ± 0.13	0.53 ± 0.49	0.47 ± 0.43	0.50 ± 0.47	0.44 ± 0.44	0.41 ± 0.40	0.34 ± 0.37
	ODRL	0.72 ± 0.19	0.96 ± 0.20	0.95 ± 0.20	0.96 ± 0.20	0.44 ± 0.24	0.95 ± 0.20	0.44 ± 0.23
gpt-oss	DCPL	0.81 ± 0.14	0.69 ± 0.45	0.63 ± 0.42	0.65 ± 0.43	0.60 ± 0.44	0.56 ± 0.41	0.47 ± 0.38
	ODRL	0.58 ± 0.25	0.96 ± 0.20	0.96 ± 0.20	0.96 ± 0.20	0.36 ± 0.29	0.96 ± 0.20	0.36 ± 0.29

semantic similarity

measured on the two NL texts using
all-MiniLM-L6-v2 from HuggingFace

Results DCPL vs ODRL

deep equality of
the two FN outputs

structural equality

Model	Language	Semantic score	(a)	(b)	(c)	(d)	$a \cdot b \cdot c$	$a \cdot b \cdot c \cdot d$
gpt5-mini	DCPL	0.82 ± 0.12	0.92 ± 0.25	0.76 ± 0.26	0.86 ± 0.27	0.83 ± 0.27	0.68 ± 0.27	0.60 ± 0.28
	ODRL	0.64 ± 0.20	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.49 ± 0.35	1.0 ± 0.0	0.49 ± 0.35
gpt4.1	DCPL	0.83 ± 0.14	0.74 ± 0.44	0.62 ± 0.40	0.67 ± 0.42	0.67 ± 0.42	0.56 ± 0.38	0.50 ± 0.36
	ODRL	0.75 ± 0.17	0.97 ± 0.16	0.94 ± 0.18	0.97 ± 0.16	0.51 ± 0.29	0.94 ± 0.18	0.49 ± 0.29
gpt4.1-mini	DCPL	0.78 ± 0.13	0.53 ± 0.49	0.47 ± 0.43	0.50 ± 0.47	0.44 ± 0.44	0.41 ± 0.40	0.34 ± 0.37
	ODRL	0.72 ± 0.19	0.96 ± 0.20	0.95 ± 0.20	0.96 ± 0.20	0.44 ± 0.24	0.95 ± 0.20	0.44 ± 0.23
gpt-oss	DCPL	0.81 ± 0.14	0.69 ± 0.45	0.63 ± 0.42	0.65 ± 0.43	0.60 ± 0.44	0.56 ± 0.41	0.47 ± 0.38
	ODRL	0.58 ± 0.25	0.96 ± 0.20	0.96 ± 0.20	0.96 ± 0.20	0.36 ± 0.29	0.96 ± 0.20	0.36 ± 0.29

semantic similarity

measured on the two NL texts using
all-MiniLM-L6-v2 from HuggingFace

Modelling fitness do vary!

Results ablation study

(increase in accuracy. baseline: no further input)

deep equality of
the two FN outputs

structural equality

DCPL

ODRL

Syntax	Few shot	Semantic score	(a)	(b)	(c)	(d)	a · b · c	a · b · c · d
present	absent	-0.04 ± 0.11	0.22 ± 0.54	0.23 ± 0.44	0.24 ± 0.52	0.30 ± 0.47	0.25 ± 0.41	0.30 ± 0.35
absent	present	0.05 ± 0.10	-0.56 ± 0.59	-0.43 ± 0.51	-0.49 ± 0.57	-0.44 ± 0.53	-0.38 ± 0.45	-0.30 ± 0.40
present	present	0.05 ± 0.09	0.23 ± 0.50	0.22 ± 0.42	0.23 ± 0.46	0.28 ± 0.44	0.20 ± 0.34	0.23 ± 0.30
present	absent	-0.20 ± 0.16	0.08 ± 0.34	0.13 ± 0.34	0.08 ± 0.34	0.35 ± 0.50	0.13 ± 0.34	0.41 ± 0.46
absent	present	-0.01 ± 0.19	0.10 ± 0.30	0.14 ± 0.34	0.10 ± 0.30	0.14 ± 0.40	0.14 ± 0.34	0.18 ± 0.34
present	present	-0.06 ± 0.18	0.10 ± 0.30	0.16 ± 0.34	0.10 ± 0.30	0.27 ± 0.48	0.16 ± 0.34	0.32 ± 0.40

semantic similarity

measured on the two NL texts using

all-MiniLM-L6-v2 from HuggingFace

Results ablation study

(increase in accuracy. baseline: no further input)

deep equality of
the two FN outputs

structural equality

DCPL

ODRL

Syntax	Few shot	Semantic score	(a)	(b)	(c)	(d)	a · b · c	a · b · c · d
present	absent	-0.04 ± 0.11	0.22 ± 0.54	0.23 ± 0.44	0.24 ± 0.52	0.30 ± 0.47	0.25 ± 0.41	0.30 ± 0.35
absent	present	0.05 ± 0.10	-0.56 ± 0.59	-0.43 ± 0.51	-0.49 ± 0.57	-0.44 ± 0.53	-0.38 ± 0.45	-0.30 ± 0.40
present	present	0.05 ± 0.09	0.23 ± 0.50	0.22 ± 0.42	0.23 ± 0.46	0.28 ± 0.44	0.20 ± 0.34	0.23 ± 0.30
present	absent	-0.20 ± 0.16	0.08 ± 0.34	0.13 ± 0.34	0.08 ± 0.34	0.35 ± 0.50	0.13 ± 0.34	0.41 ± 0.46
absent	present	-0.01 ± 0.19	0.10 ± 0.30	0.14 ± 0.34	0.10 ± 0.30	0.14 ± 0.40	0.14 ± 0.34	0.18 ± 0.34
present	present	-0.06 ± 0.18	0.10 ± 0.30	0.16 ± 0.34	0.10 ± 0.30	0.27 ± 0.48	0.16 ± 0.34	0.32 ± 0.40

semantic similarity

measured on the two NL texts using

all-MiniLM-L6-v2 from HuggingFace

Exposure to schema has a negative impact to semantic score, but for DCPL few-shots compensate.

Exposure to schema greatly improves the symbolic output scores, but few-shot only degrades it.

Preliminary conclusions

- Initial study meant to prove the experimental pipeline.
- Data and pipeline available at <https://github.com/uva-cci/nl12fr-2025-neurosymbolic>
- So far simple languages (only schema), and simple scenarios. We will expand on this (if you are working on normative languages, you may receive an email from us!)
- Yet, we already observed an interesting ***trade-off between formal consistency and domain coherence***: best performance on deep equality coincides with the worst performance on the semantic score, and vice-versa.

Testing Modelling Fitness of Normative Specification Languages for Large Language Models

NLL2FR Workshop, joint with JURIX 2025,
Torino, 9 December, 2025



Giovanni Sileno
g.sileno@uva.nl

Andrea Marino
a.marino@uva.nl

University of Amsterdam