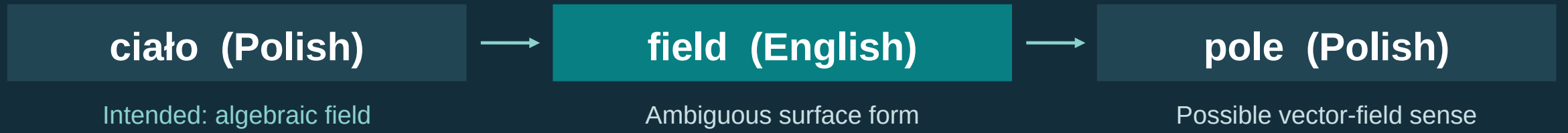
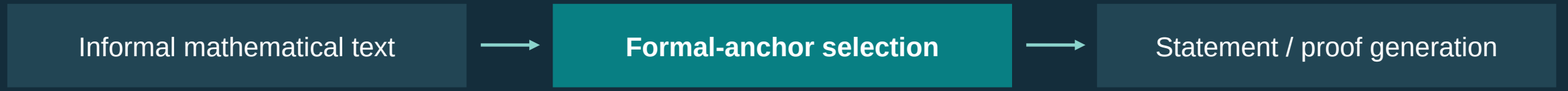
FACBench

A Benchmark for Formal-Anchor Collisions in Multilingual Mathematical Grounding

MOTIVATION: TRANSLATION CAN CHANGE THE INFERRED CONCEPT



Illustrative drift when context is lost — not a measured translation result.



FACBench audits this choice

Piotr Kuterba · Józef Spałek

Jagiellonian University, Kraków, Poland

Five languages — one intended concept

Different linguistic forms should preserve the intended mathematical role.

52
candidate concepts

130
contrastive statements

5
languages: EN / PL / DE / FR / ES

13
bidirectional families

01 CURATE MEANING

Semantic concept cards

45 verified Mathlib mappings
+ 7 proxy mappings

→

02 RANK THE INVENTORY

Frozen retrievers

Same 52 candidates
for every query

→

03 AUDIT THE CHOICE

Acc@1 + Recall@5

Prespecified-confusable
selection among errors

Family-level uncertainty

13 pairs × 2 directions × 5 languages = 130 queries

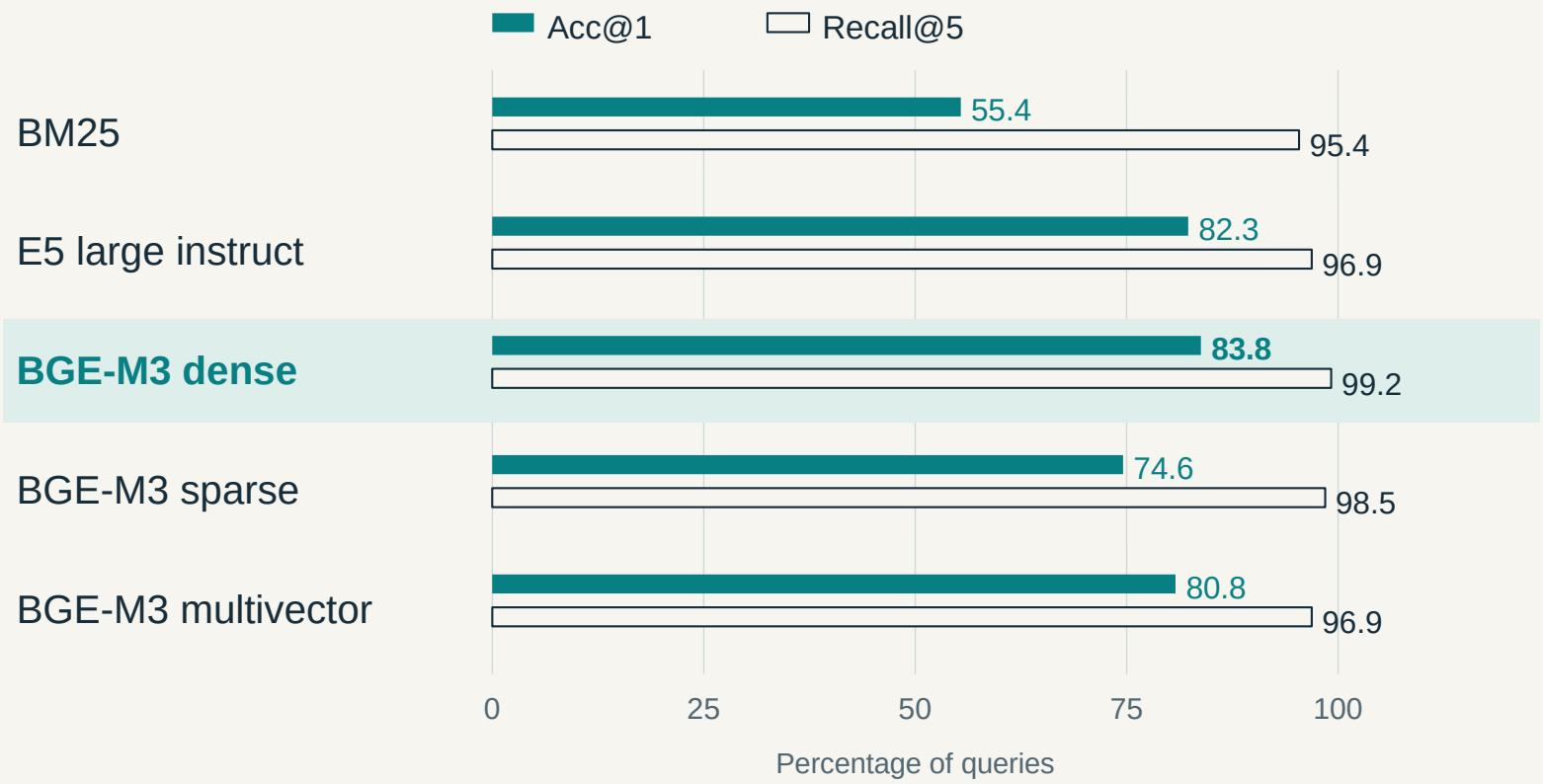
Primary endpoint: concept-ID

Not accuracy over 52 independently verified declarations.

BM25: word matching · E5: vector similarity · BGE-M3: dense / sparse / multivector

Correct target available. Wrong concept selected.

Final adjudicated snapshot · Semantic-only cards · Frozen off-the-shelf retrievers



130 statements · 13 source families · 52 candidates

Semantic-only cards: English label + definition, five-language aliases, no formal-anchor metadata.

BGE-M3 DENSE

99.2%

Recall@5 / in the first five

83.8%

Acc@1 / correct first choice

109 first · 20 at ranks 2–5

1 target outside the top five

Not every near miss is interchangeable

BGE-M3 DENSE / FINAL SNAPSHOT

12 / 21

Top-1 errors select the prespecified confusable.



12 confusable — 9 other

ILLUSTRATIVE REQUEST / NOT A SCORED QUERY

A basis in linear algebra, not a basis of open sets.

MATCHES THIS ROLE

`Module.Basis`

Linear-algebraic basis

DIFFERENT FORMAL ROLE

`TopologicalSpace.IsTopologicalBasis`

Topological basis

Shared surfaces: basis / base / baza

Both are valid objects. They are not interchangeable for this request.

The guard helps its designed regime, not every query

BGE-M3 dense + source/context ranking · Reciprocal rank fusion (RRF), k = 60

Controlled queries

61.2% → **87.6%**

Retriever alone → After fixed fusion

+26.4 pp

Contrastive queries

83.1% → **73.1%**

Retriever alone → After fixed fusion

−10.0 pp

Separate guard-fusion experiment; dataset version differs from slide 3.

The same reversal occurs for E5 and BGE-M3 multivector.

Preserve the mathematical concept, not just a relevant match.

01

Measure selection

Acc@1 alongside Recall@5.

02

Inspect confusables

Known role conflicts, not all errors.

03

Keep the scope explicit

Targeted tests, not prevalence.

The correct concept can be available, while a non-interchangeable neighbour ranks first.

Closed diagnostic only. Downstream correction or propagation remains untested.

ACKNOWLEDGEMENTS

National Science Centre, Poland (NCN)

UMO-2021/41/B/ST3/04070 · 2023/49/B/ST3/03545

Piotr Kuterba · Józef Spałek

Jagiellonian University, Kraków, Poland

Data + code: doi.org/10.57967/hf/11400