Towards a Danish Semantic Reasoning Benchmark

- with lexical-semantic resources as a gold standard

Bolette S. Pedersen, Nathalie Sørensen, Sussi Olsen, Sanni Nimb, Simon Gray

DET DANSKE SPROG- OG LITTERATURSELSKAB

UNIVERSITY OF COPENHAGEN

Can lexical-semantic resources be repurposed?

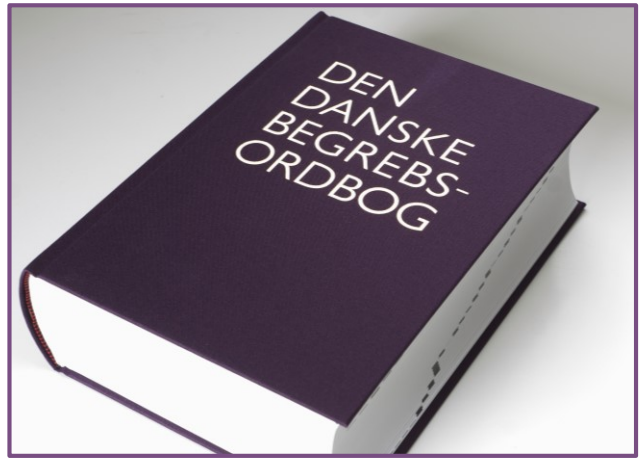
Aim

To explore whether our lexical-semantic resources can be meaningfully applied as a **gold standard** for semantic reasoning testsets for Danish

1) Can the datasets be compiled (semi)automatically from the resources?
   a) How much manual interference and curation is needed?
2) To which degree will such datasets actually challenge the ceiling performance of state-of-the-art LLMs?
3) Which reasoning aspects in our resources are harder for the models to assess? (concrete vs abstract, inference task vs relatedness tasks etc.)

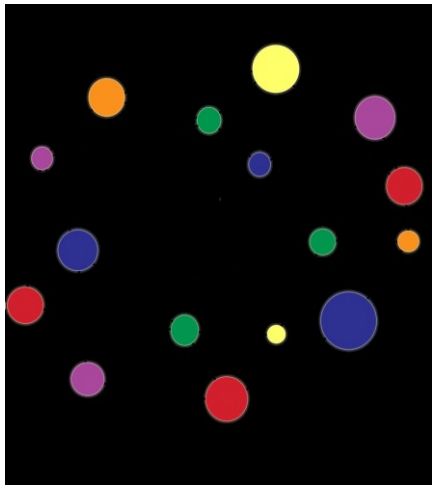
Which lexical-semantic resources?

- **The Danish wordnet**, DanNet (Pedersen et al., 2009), which relates all concepts to an ontology and encodes an internal taxonomy of hyponymy together with an additional number of semantic relations describing the core meaning components of a given concept
- **The Danish Thesaurus** (Nimb et al., 2014) where the ordering of the words in chapters, sections and subgroups depending on their topic and semantic relatedness can be used to deduce semantic similarity and synonymy



Which lexical-semantic resources

- **The Danish FrameNet Lexicon** (Nimb et al., 2017; Nimb, 2018) where all verbs and deverbal nouns are assigned a reference to the semantic frames inventory from **Berkeley FrameNet** (Baker et al., 1998) thus, enabling the extraction of e.g. change-of-state verbs, communication and mental verbs.



- **The Central Word Register for Danish** (Nimb et al., 2022; Pedersen et al., 2022) a recently developed computationa lexicon for Danish with a simplified (i.e. coarse-graine sense inventory. COR.SEM (the semantic module) contains usage examples and sentiment.

COR.SEM
DET CENTRALE ORDREGISTER SEMANTISK MODUL

OUR PILOT STUDY:
Seven semantic tasks
across 26 datasets

Prompting ChatGPT 3.5 Turbo, 4 and 4o

To evaluate whether our datasets are actually challenging enough for some of the most recent LLMs, we prompt ChatGPT 3.5 turbo, ChatGPT 4 and ChatGPT 4o with our test instances

Each task has its own prompt description

The model is asked to answers either by true / false or by one or two words (to ease the evaluation)

Task 1 (from DanNet)

Inference tests: The ability to infer the ontological status and other core meaning components of a given concept in context.

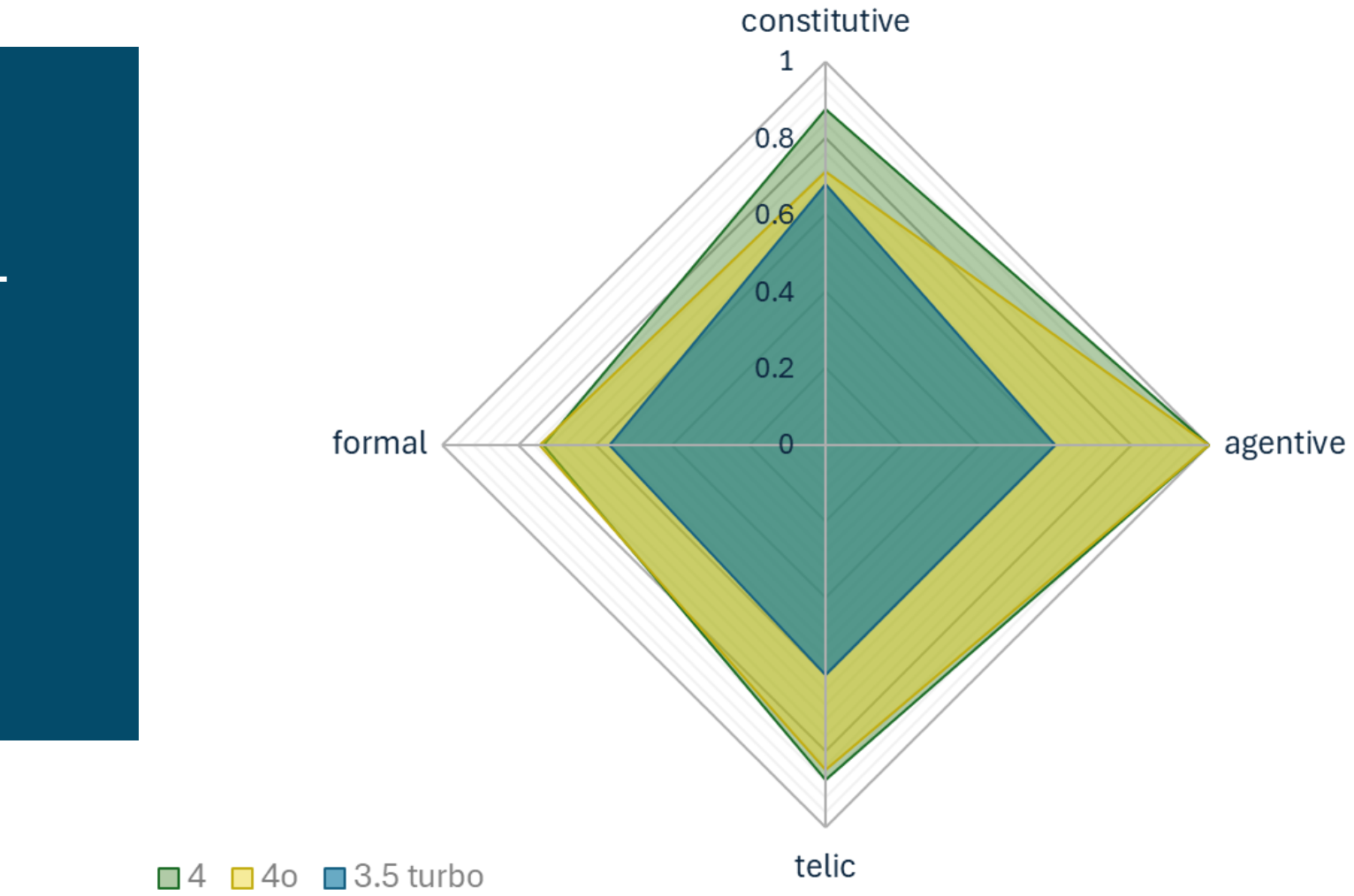
- DanNet contains approx. 350,000 semantic relations across 70,000 concepts from an inventory of 15 different relation types
- Following the theory of Pustejovsky's Generative Lexicon (Pustejovsky, 1995), the relations are organised into four so-called qualia roles relating to a concept's:
  1. **Ontological type** (formal role)
  2. **Coming about** (agentive role, origin),
  3. **Function** (telic role),
  4. **Parts and whole and other characteristics** (constitutive role)

Generic templates are generated for each role

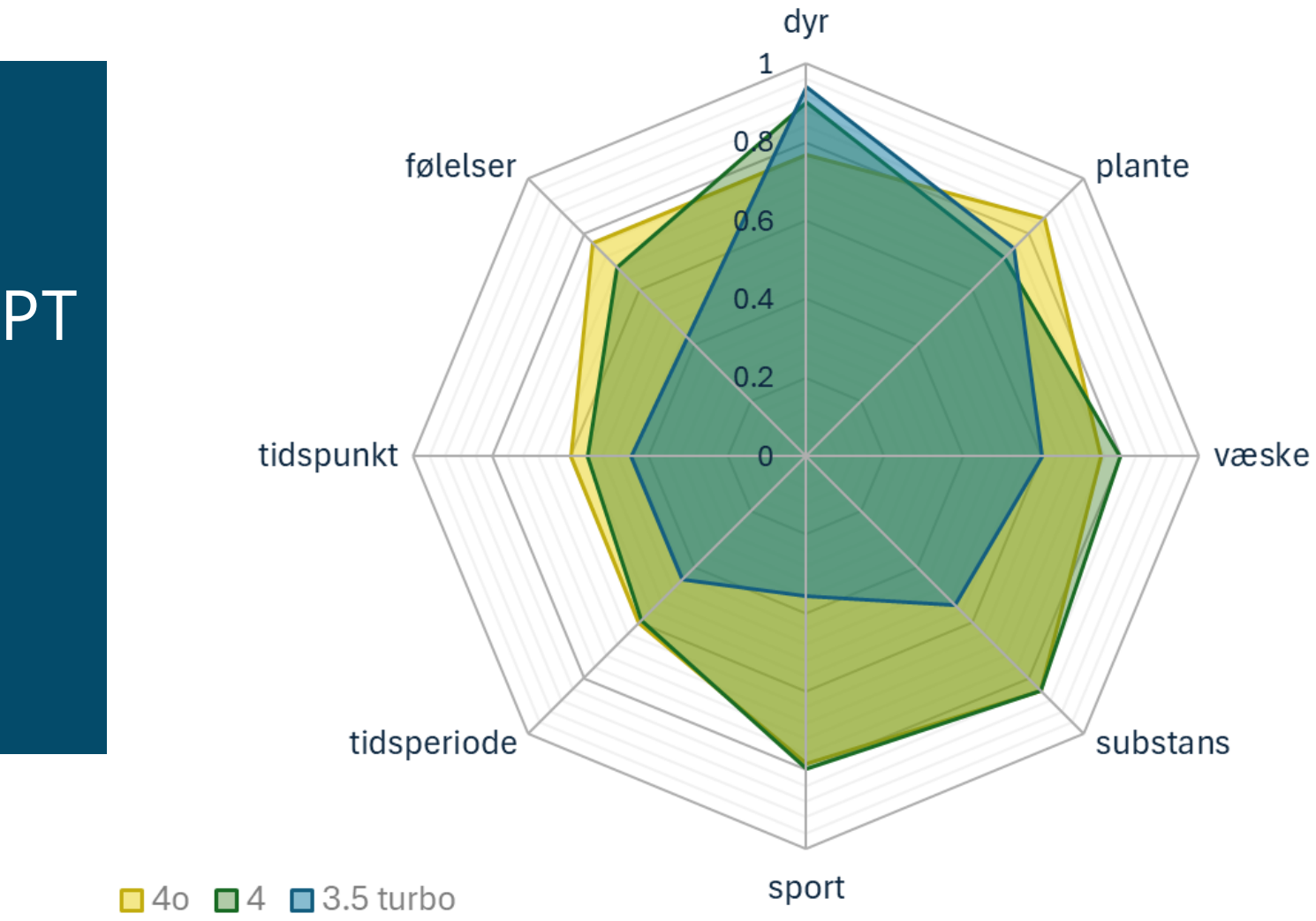
Task 1: Ontological status and core meaning

Qualia and Ontotype	Generated utterances	Translation
FORMAL Feeling; Creature	P: *Sympati er en følelse; Tryghed er en følelse* Q: *Spøgelse er en følelse* (false)	P: Sympathy is a feeling; Safety is a feeling Q: Ghost is a feeling (false)
FORMAL Liquid; Quantity	P: *En bouillon er en væske; En drik er en væske* Q: *En slurk er ikke en væske* (true)	P: A broth is a liquid; A drink is a liquid Q: A sip is not a liquid (true)
AGENTIVE Semiotic; Artifact	P: *Man laver en roman ved at skrive den; Man laver et essay ved at skrive det* Q: *Man laver ikke en hat ved at skrive den* (true)	P: You make a novel by writing it; You make an essay by writing it Q: You don't make a hat by writing it (true)
AGENTIVE Food; Liquid	P. *Man laver et tog ved at fremstille det; Man laver en ret ved at tilberede den* Q: *Man laver te ved at porchere den* (false)	P. You make a train by manufacturing it; You make a dish by cooking it Q: You make tea by poaching it (false)
TELIC Garment; Artifact	P: *Man tager en frakke på for at holde sig varm; Man tager en hue på for at holde sig varm* Q: *Man tager en ring på for at holde sig varm* (false)	P: You put on a coat to keep warm; You wear a hat to keep warm Q: You wear a ring to keep warm (false)
TELIC Instrument	P: *Man bruger en kniv til at skære med; Man bruger en hammer til at hamre med* Q: *Man bruger ikke et rivejern til at rive med* (false)	P: You use a knife to cut with; You use a hammer to hammer with Q: You don't use a grater to grate with (false)
CONSTITUTIVE BodyPart; Part	P: *En hånd kan ikke have et øje; Et ansigt kan have en mund* Q: *Et fly kan have en propel* (true)	P: A hand cannot have an eye; A face can have a mouth Q: A plane can have a propeller (true)

How well does ChatGPT perform on the qualia roles?



How well does ChatGPT perform on different ontological types?



Task 2: Entailments from events

To which extent do LLMs infer the result of an event?

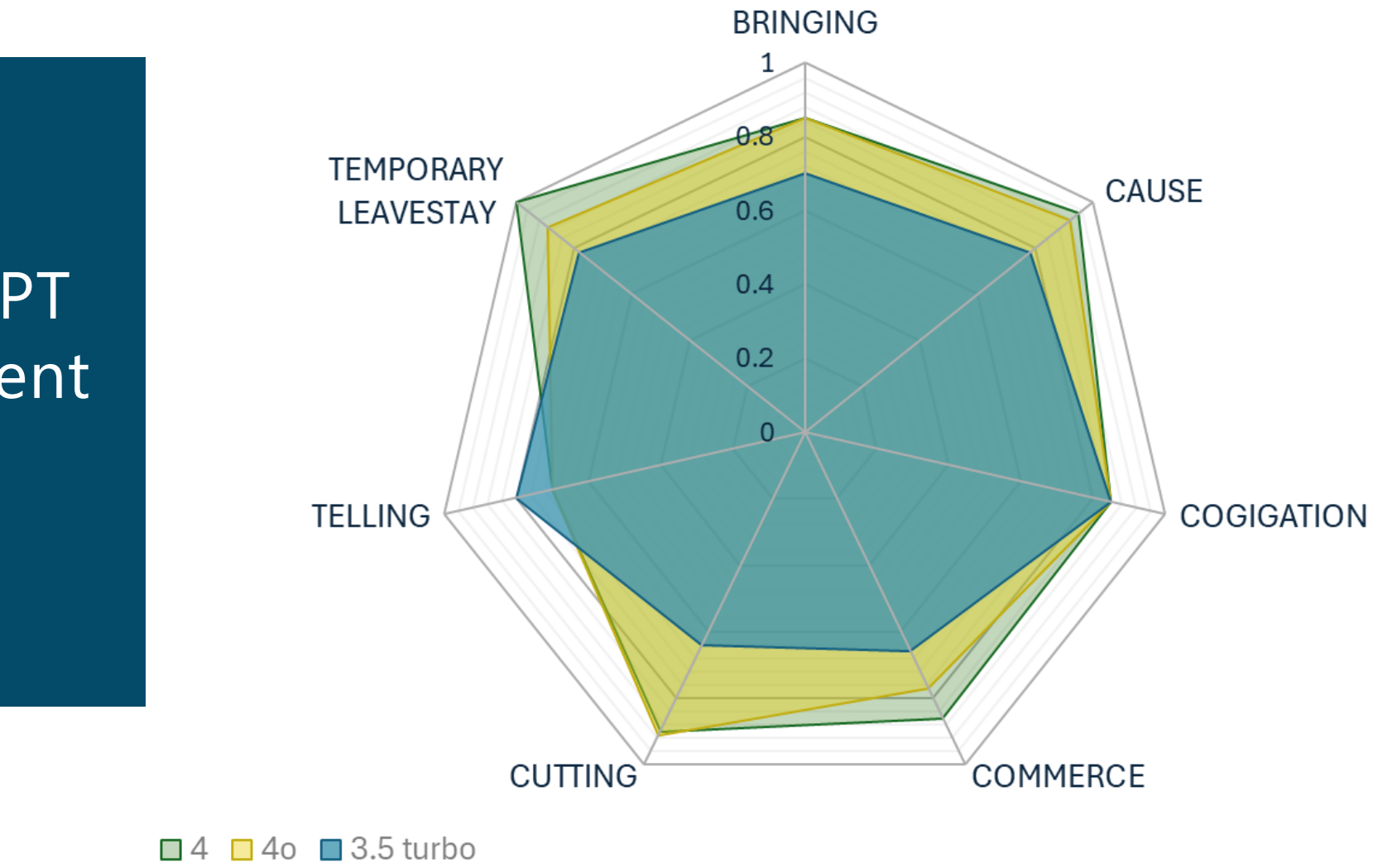
- Danish FrameNet contains 671 different frames assigned to 5,300 Danish verbs and 6,490 deverbal nouns, and refers to the semantic event ontology of Berkeley FrameNet's (Baker et al., 1998) where it is spelled out what kind of event a frame refers to, what kind of result is achieved (if any), and which frame elements are typically evoked.

For each of the tested frames, we design one or a few generic templates, like for BUYING Peter X bogen af/fra Y ('Peter X the book from Y')

Task 2: Entailment from FrameNet

Semantic frame	Generated utterance	Translation
BRINGING	P: *Peter bringer mad ud til Pia.* *Q: Pia har nu mad.* (true)	P: Peter brings out food to Pia. Q: Pia now has food (true)
CAUSE	P: *Eksplosionen medførte svære skader på bygningen.* Q: *Efter eksplosionen var bygningen ubeskadiget.* (false)	P: The explosion resulted in severe damages on the building Q: After the explosion the building was un-damaged (false)
BUYING	P: *Peter købte bogen af Anne* Q: *Nu ejer Anne bogen* (false)	P: Peter bought the book from Anne, Q: Now Anne owns the book (false)
CUTTING	P: *Pia klipper rebet over.* Q: *Pia har nu to kortere reb.* (true)	P: Pia cuts the robe, Q: Pia now has two shorter robes (true)
TELLING	P: *Peter fortalte Pia om forlovelsen*, Q: *Pia kender nu til forlovelsen* (true)	P: Peter told Pia about the engagement, Q: Pia now knows about the engagement (true)

How well does ChatGPT perform on the different types of events?



Task 3: Synonymy selection from BEO

The Danish Thesaurus (BEO) contains 22 chapters and 888 sections with more than 100,000 lemmas and 130,000 word senses divided into groups with up to three levels of semantic similarity and relatedness marked in the structure

How well do the models identify synonymous words?

We give the model a target and four candidates:

1. a correct synonym
2. a similar word (same section)
3. a related word (same chapter)
4. a random word (different chapter)

Task 4 and 5: Semantic similarity and relatednes from BEO

We take inspiration from the dasem word intrusion dataset (Nielsen and Hansen, 2017)

Word intrusion: list of words where one is an outlier. The rest are semantically similar or related. The task is to identify the outlier.

We create different granularities of the dataset, where the outlier is more or less related to the rest of the group.

Similar terms: *øluft* (island air), *havluft* (sea air), *havgus* (seagull),

Related but not similar term (semantic outlier): *øhav* (archipelago)

Task 6: WSD WiC(Word-in-Context) from COR.SEM

We take advantage of the coarse-grainedness of COR.SEM with examples to each sense to generate a WiC Corpus.

We take a words usage examples and pair them according to two patterns:

1. Pair examples of same COR.SEM sense
2. Pair examples of two different COR.SEM senses

The question is whether

The example pair come from one or two senses:

*Vi passerede skibene i en **afstand** af ca. 40 meter*
We passed the ships in a **distance** of about 40 meters

*brugen af fagsprog skaber en **afstand** mellem læge og patient*
The use of technical language creates a **distance** between the doctor and the patient

Task 7: Sentiment from COR.SEM

COR.SEM also contains sentiment values for many senses.

We create a sentiment-in-context dataset by taking the usage examples and sentiment values.

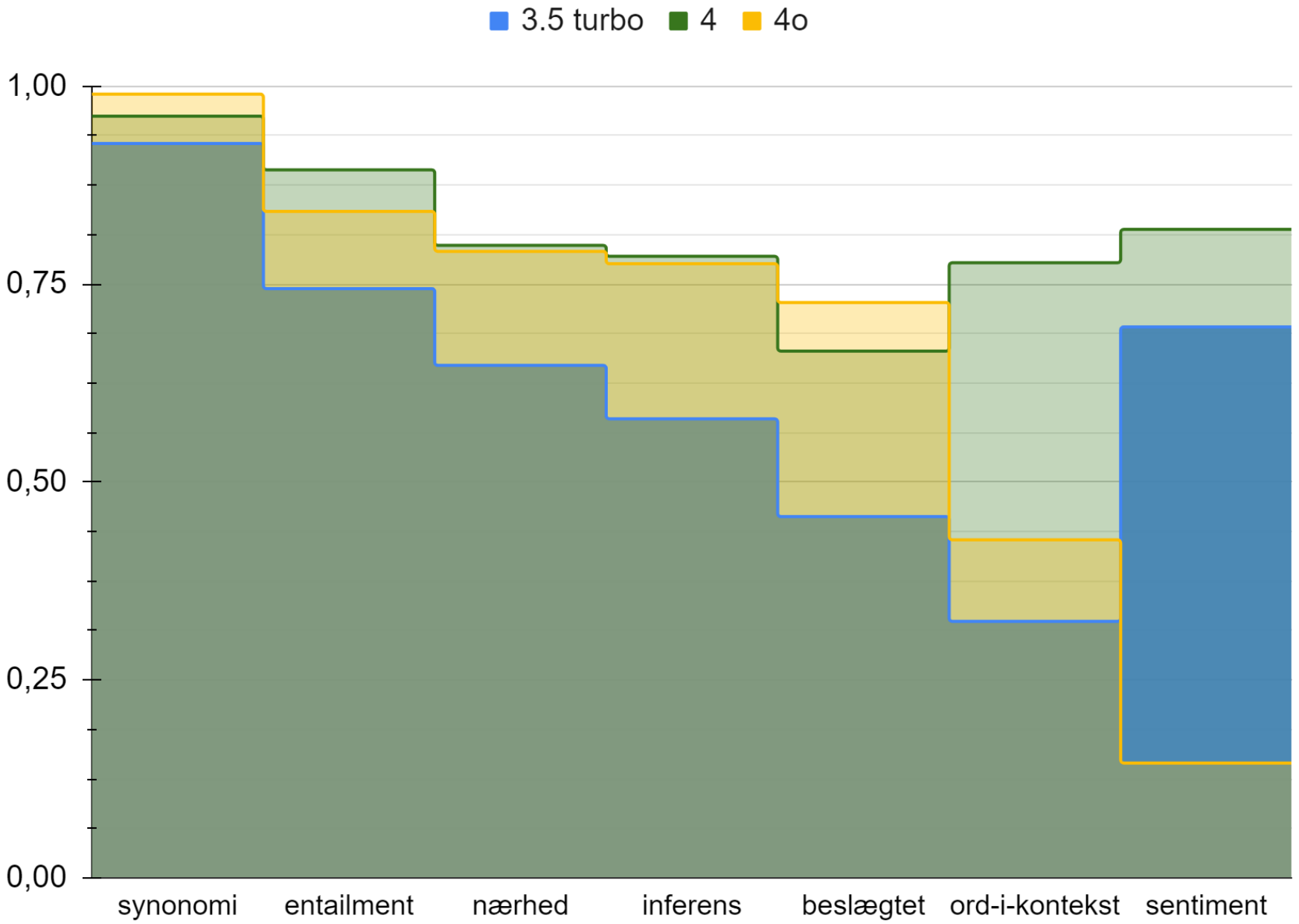
The question is

how negative (-3) or positive (+3) a word is in the specific context:

*Mandens tilværelse er **knust**. Han sidder bare i en stol og ser apatisk frem for sig.*
The man's existence is **shattered**. He just sits in a chair, staring apathetically ahead.

*Hendes øjne **stråler**, og hun har det dejligt.*
Her eyes are **shining**, and she feels wonderful.

All Seven Tasks



Ongoing work

We are currently **scaling up** the datasets to include more test instances and to thereby test more broadly the mastering of vocabulary and different semantic and ontological categories.

We are also expanding **to other, harder reasoning tasks**, including the comprehension of idiomatic expressions and metaphors

We are planning to do some parallel tests with humans to avoid **'superhuman'** benchmarks

Thank you

- The project can be followed: https://cst.ku.dk/english/projects/the-benchmark-project

- Data is available: https://github.com/kuhumcst/danish-semantic-reasoning-bench-mark

Publications:

- Pedersen, B. S., Sørensen, N. C. H., Olsen, S., & Nimb, S. (2024). Evaluering af sprogforståelsen i danske sprogmodeller – med udgangspunkt i semantiske ordbøger. *NyS - Nydanske Sprogstudier*, *65*, 8-40. [1].
- Pedersen, B. S., Sørensen, N. C. H., Olsen, S., Nimb, S., & Gray, S. (2024). Towards a Danish Semantic Reasoning Benchmark - Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. I *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (s. 16356). ELRA and ICCL.