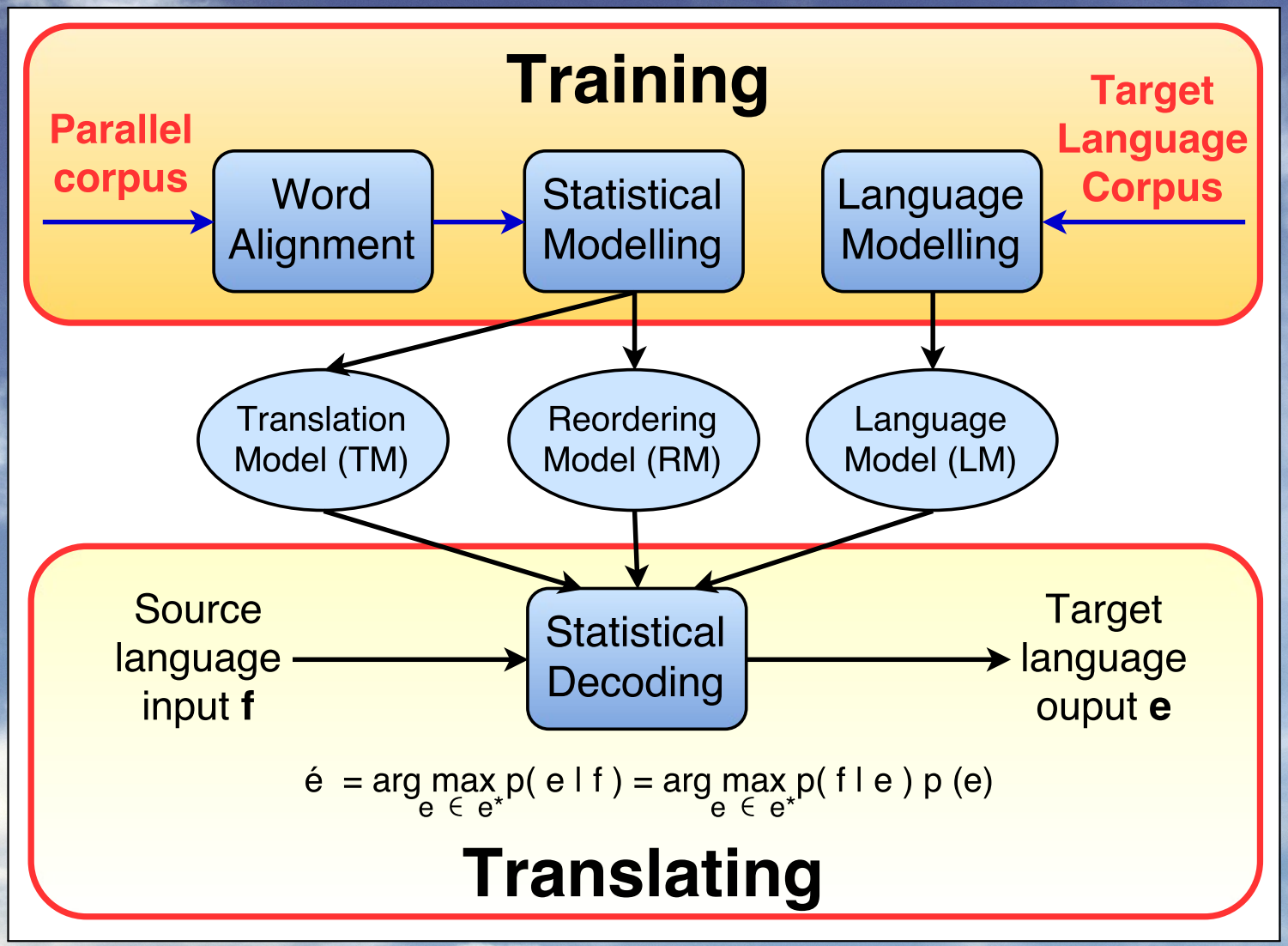
REMEDI: Robust and Efficient Machine Translation in a Distributed Infrastructure

Bi-Annual Report: Month 24

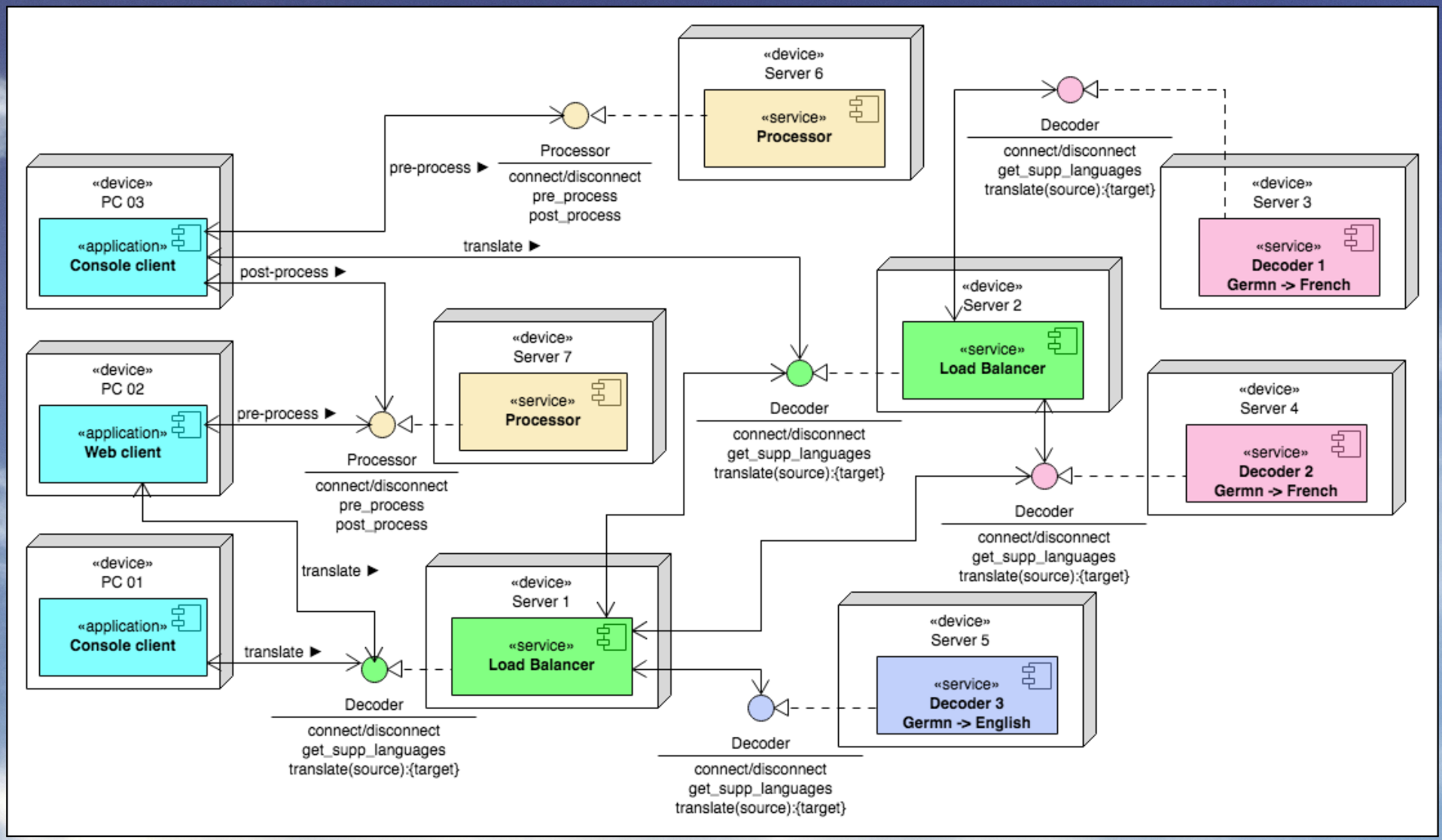
Dr. Ivan S. Zapreev • UvA • ivan.zapreev@uva.nl
Dr. Christof Monz • UvA • c.monz@uva.nl

Retrospective

Statistical Machine Translation

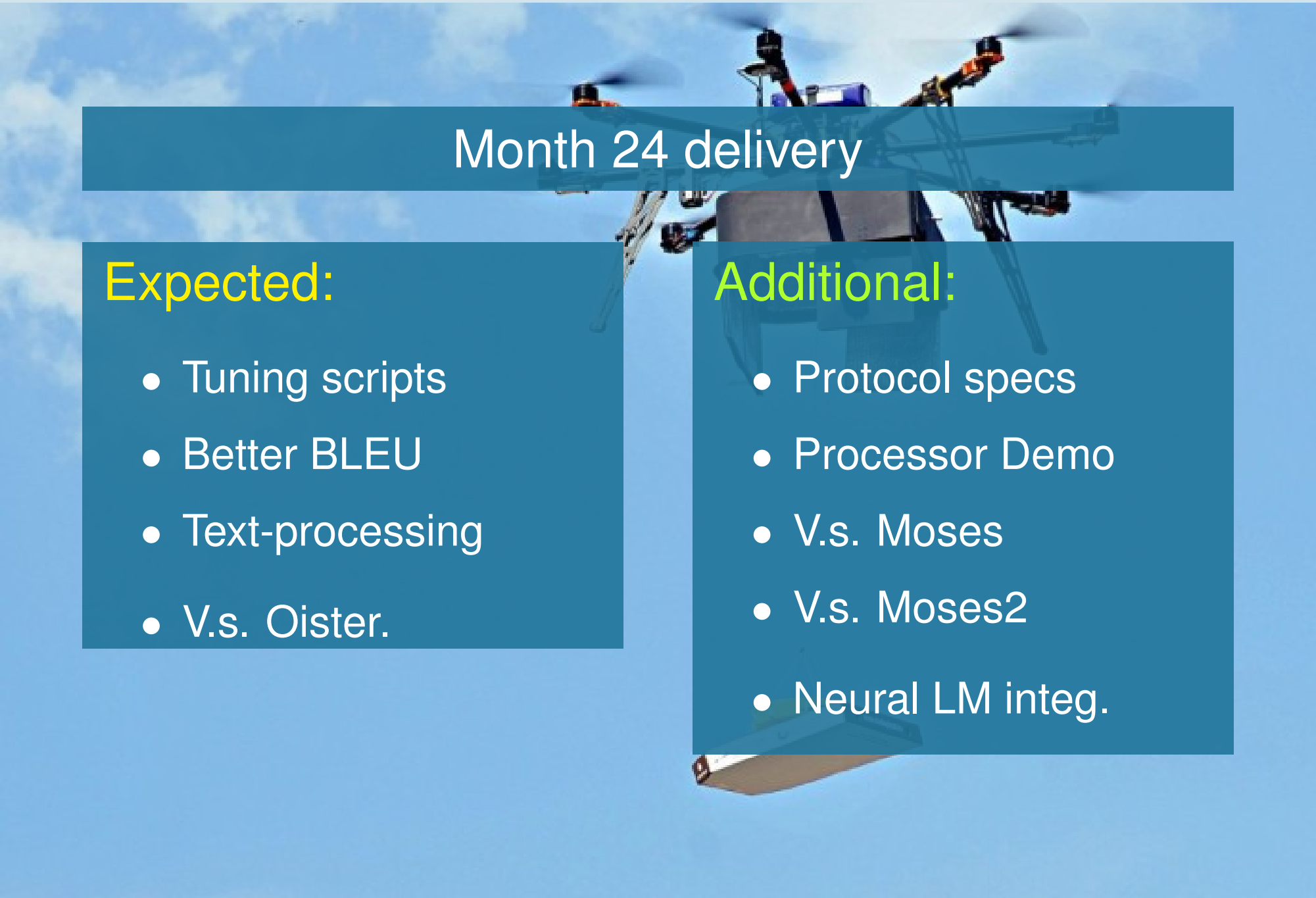


Retrospective: Previous 3 deliveries



Functionality, Performance, Profiling, Documentation

Present days



Month 24 delivery

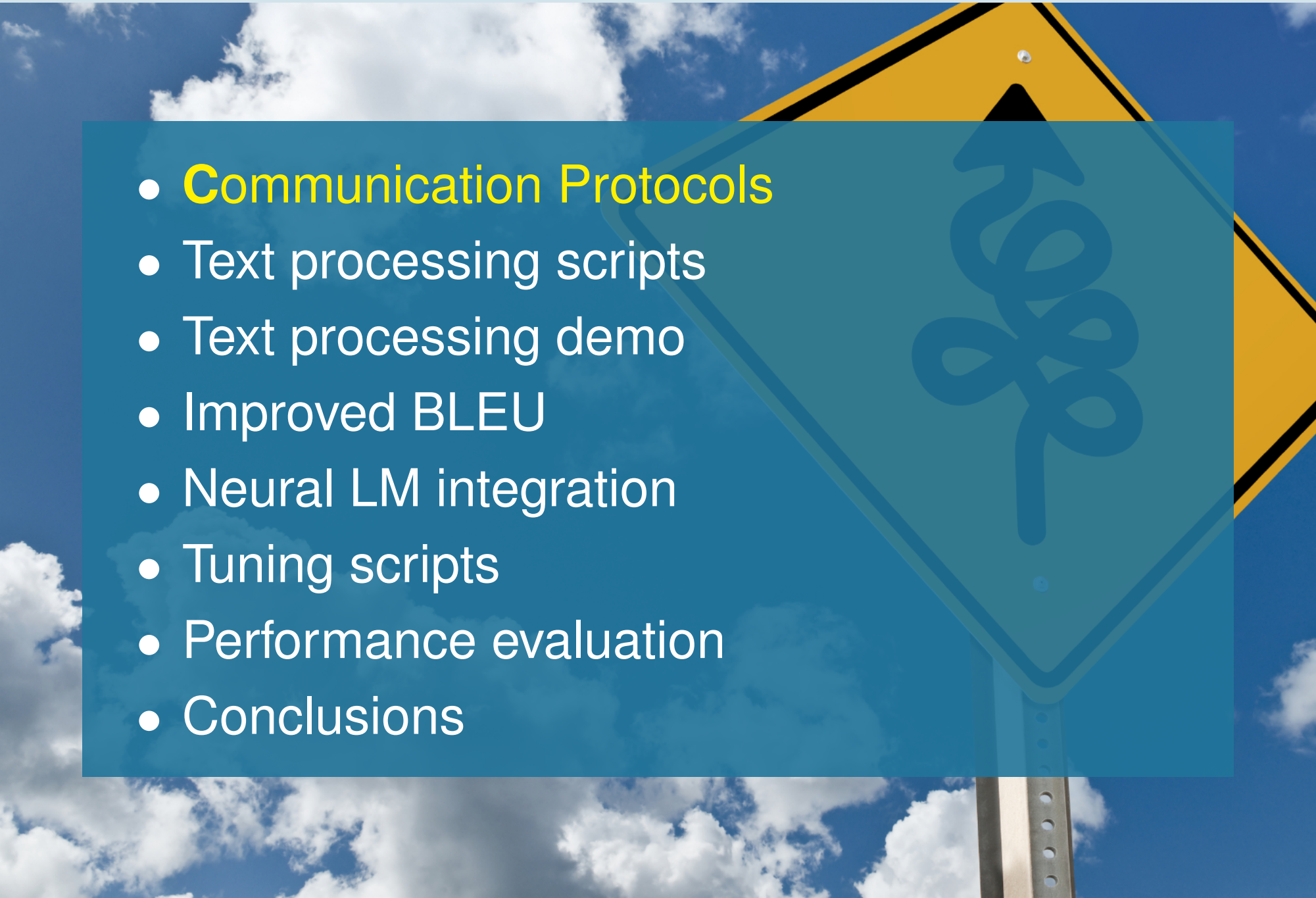
Expected:

- Tuning scripts
- Better BLEU
- Text-processing
- V.s. Oister.

Additional:

- Protocol specs
- Processor Demo
- V.s. Moses
- V.s. Moses2
- Neural LM integ.

Table of Contents

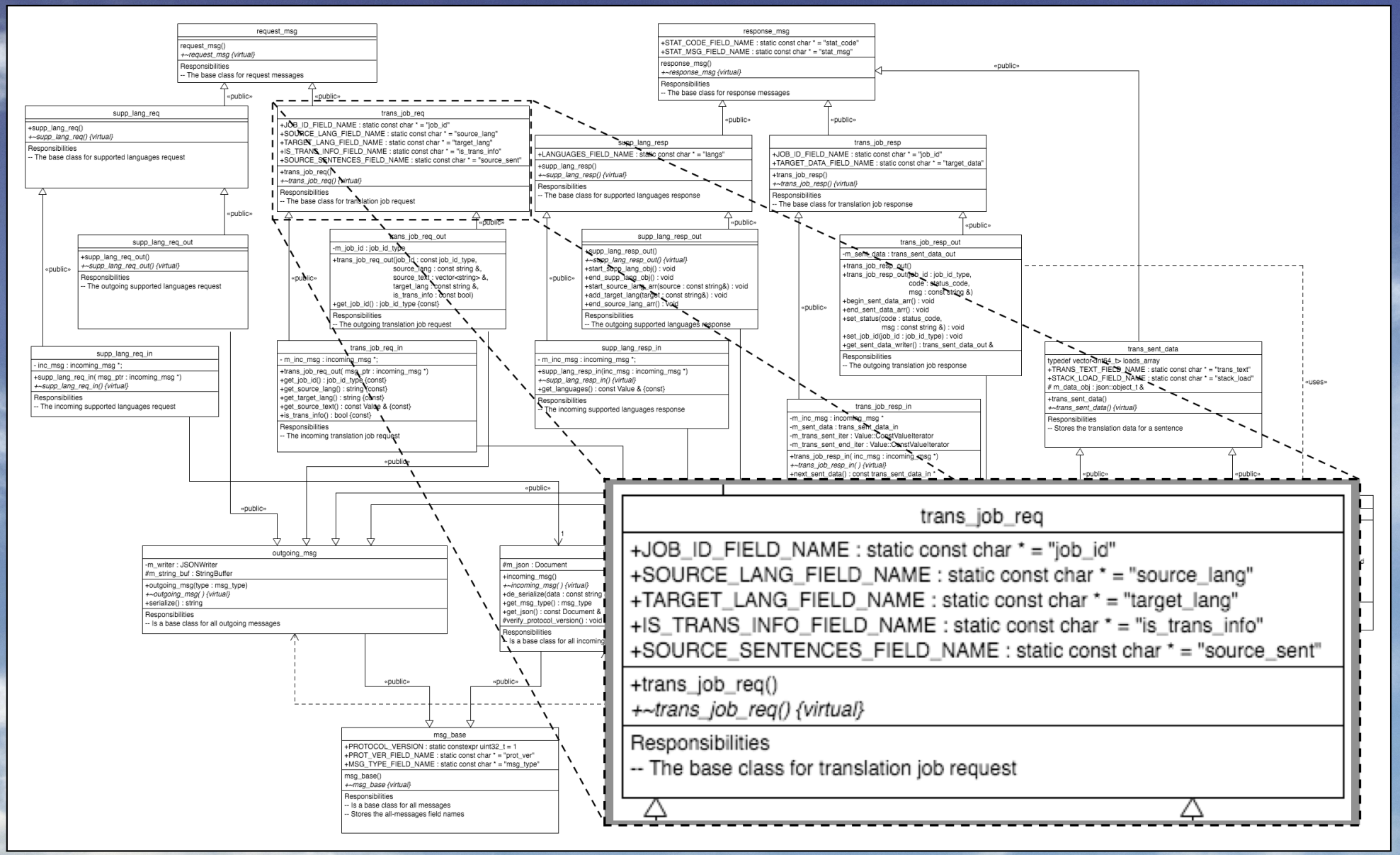


Protocols: Supported communications

- **(T)** - Translation
- **(S)** - Supported languages
- **(P)** - Pre-/Post-processing

Req/Resp	Server	Balancer	Processor
Client	(S), (T)	(S), (T)	(P)
Balancer	(S), (T)	(S), (T)	

Protocols: Class diagram



Protocols: JavaScript Object Notation

```
{
  "prot_ver" : 0,
  "msg_type" : 3,
  "job_id" : 101,
  "priority" : 10,
  "source_lang" : "german",
  "target_lang" : "english",
  "is_trans_info" : true,
  "source_sent" : [
    "how are you ?",
    "i was glad to see you .",
    "let us meet again !"
  ]
}
```

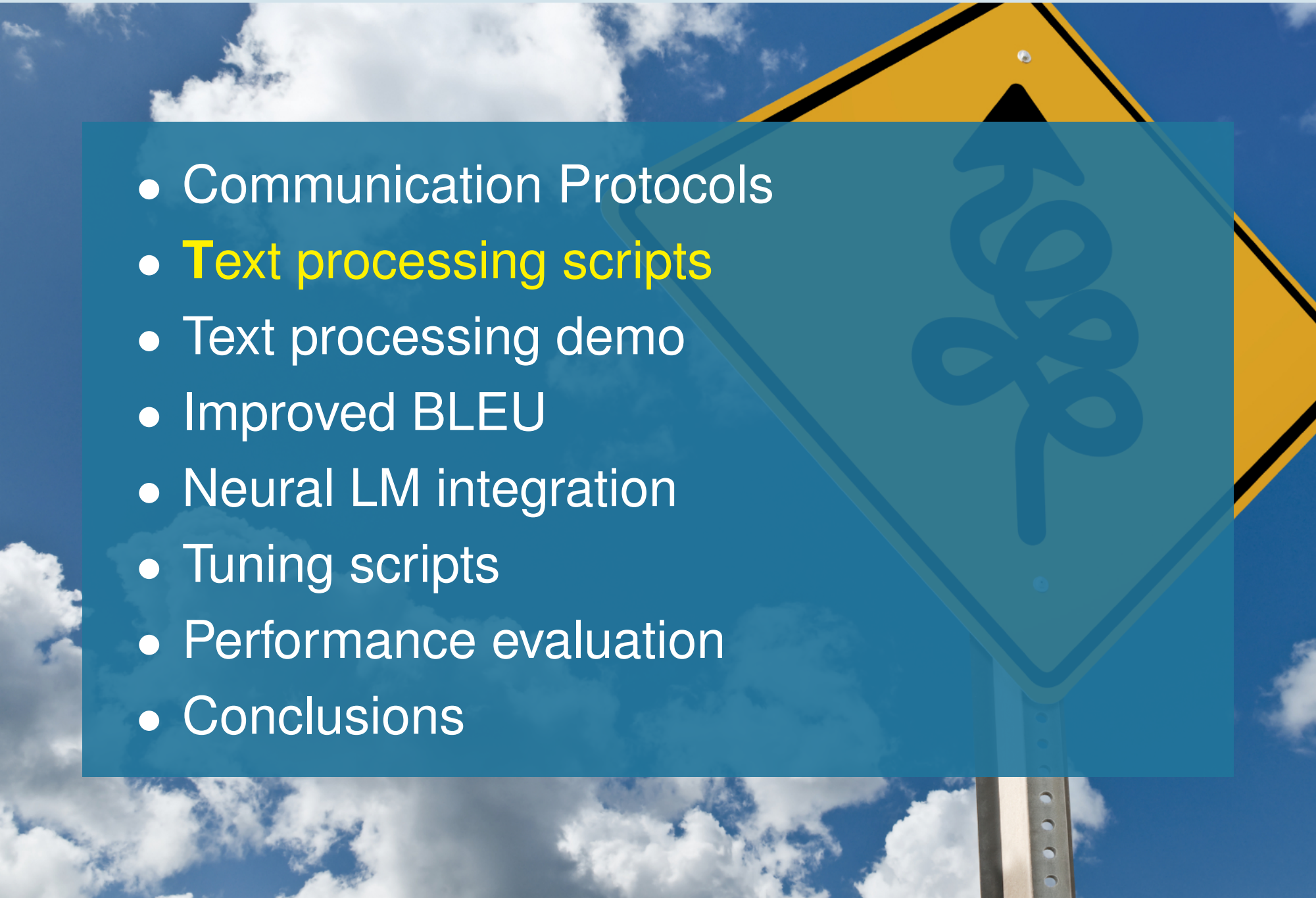
Translation Request, JSON

Protocols: JavaScript Object Notation

```
{
  "prot_ver" : 0,
  "msg_type" : 4,
  "job_id" : 101,
  "stat_code" : 3,
  "stat_msg" : "Some translation tasks were canceled",
  "target_data" : [
     {
       "stat_code" : 2,
       "stat_msg" : "OK",
       "trans_text" : "Wie geht es dir ?",
       "stack_load" : [ 3, 67, 90, 78, 40, 1 ]
     },
     {
       "stat_code" : 4,
       "stat_msg" : "The service is going down",
       "trans_text" : "i was glad to see you ."
     },
     {
       "stat_code" : 2,
       "stat_msg" : "OK",
       "trans_text" :  "Lass uns uns wiedersehen!",
       "stack_load" : [ 7, 76, 98, 90, 56, 47, 3 ]
     }
   ]
}
```

Translation Response, JSON

Table of Contents

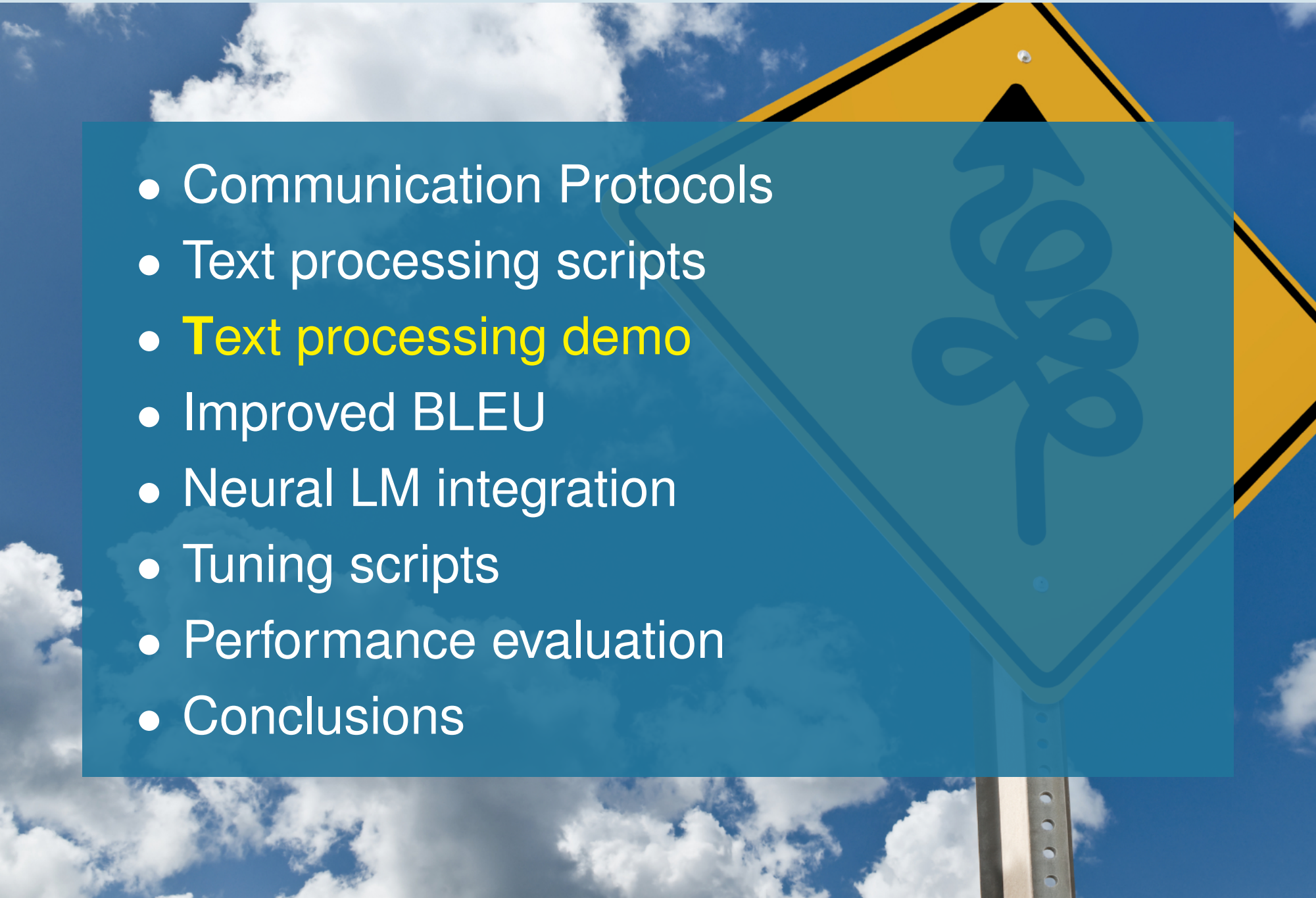


Text processing scripts

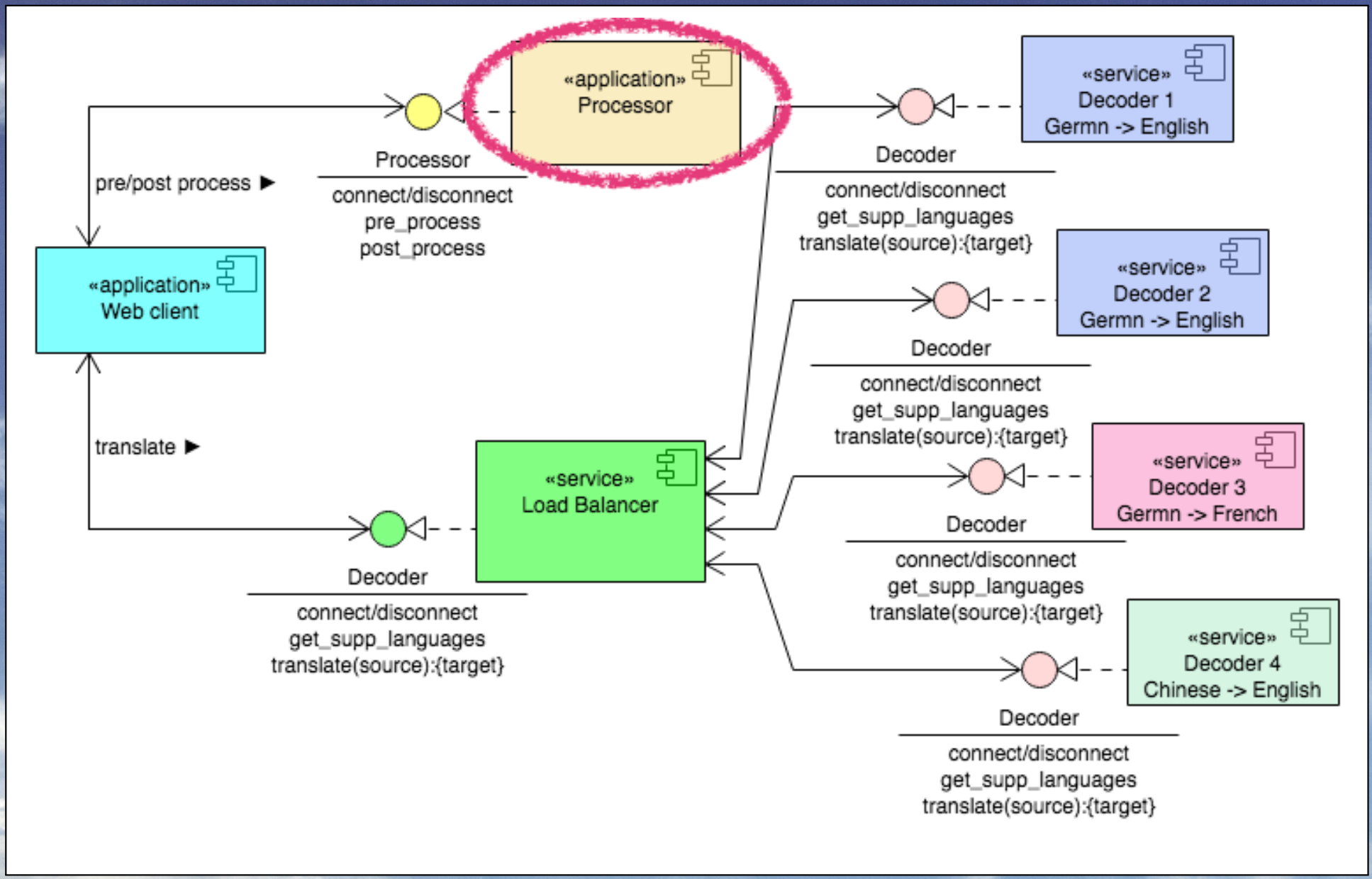
Integrated third-party tools and services

- **`pre_process_nltk.sh`:**
  - Language detection via NLTK stop words [LB02, COJA15]
  - Template generation for text structure restoration
  - Sentence splitting using NLTK and Stanford Core NLP
  - Tokenization and lowercasing via NLTK [MSB+14]
- **`post_process_nltk.sh`:**
  - Sentence capitalization as a separate option
  - Text de-tokenization via MTMonkey [AORP13]
  - True-casing via Moses [KHB+07] or Truecaser [LIRK03].
  - Text structure restoration from a pre-generated template.

Table of Contents

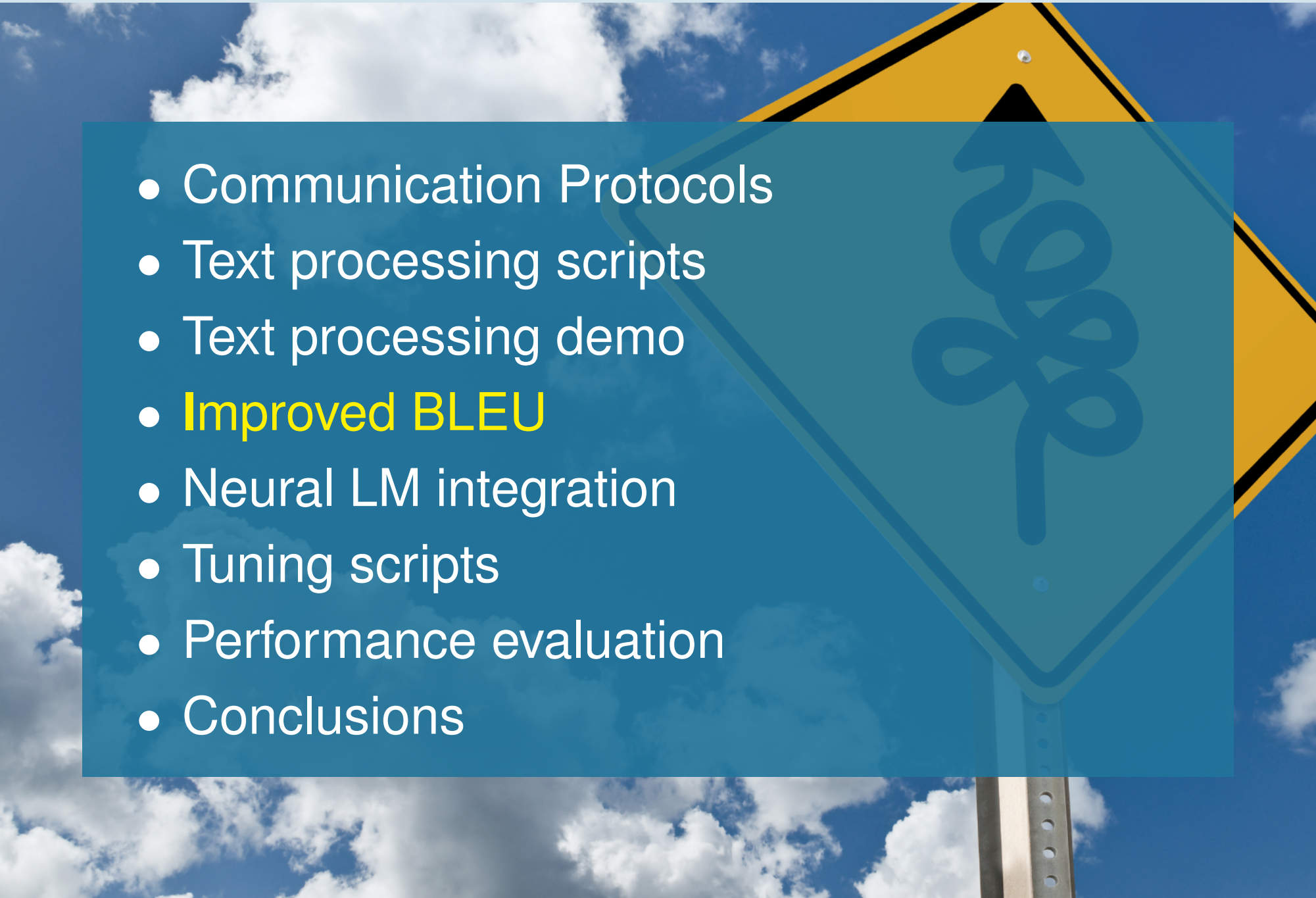


Demo: Deployment



Demo: Text processing demo

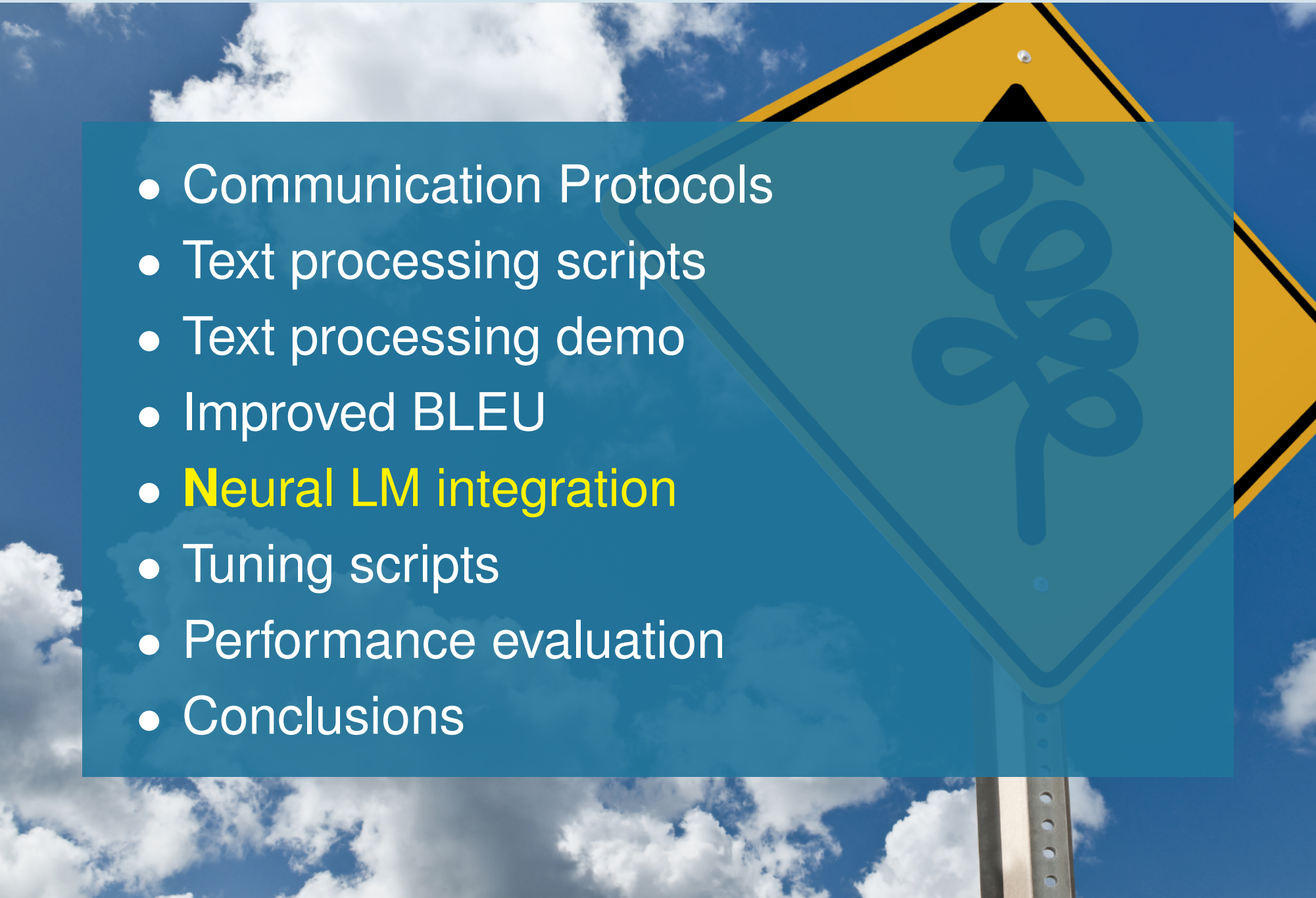
Table of Contents



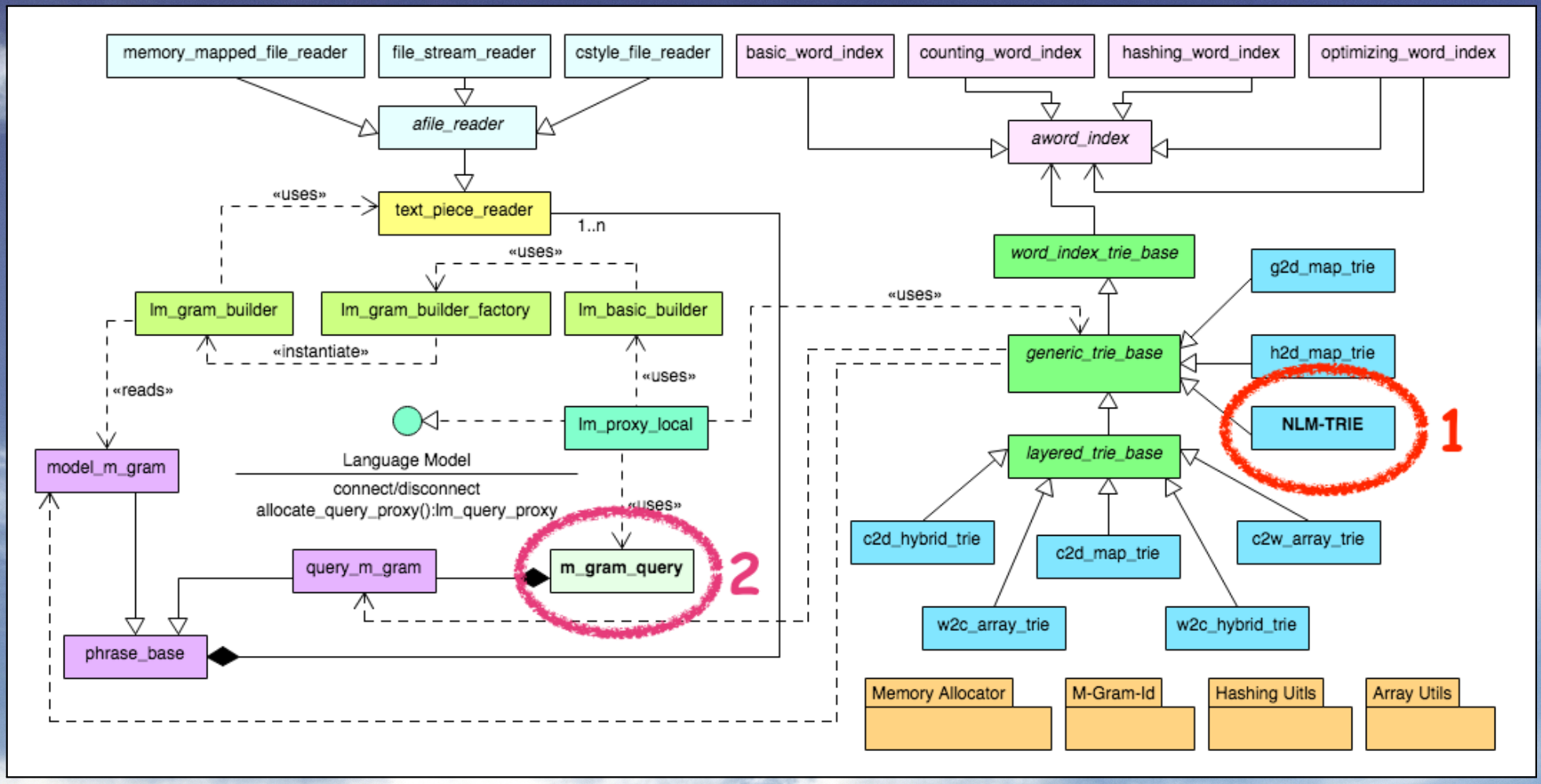
Improved BLEU

- Chinese to English:
  - **MT-**04 [Gro10]
  - CTB segmented [MSB+14]
  - 1788 sentences
  - 49582 tokens
- **Oister**: 36.8 BLEU
- REMEDI m-18: $\approx$ 23 BLEU
- REMEDI m-24: 36.72 BLEU

Table of Contents

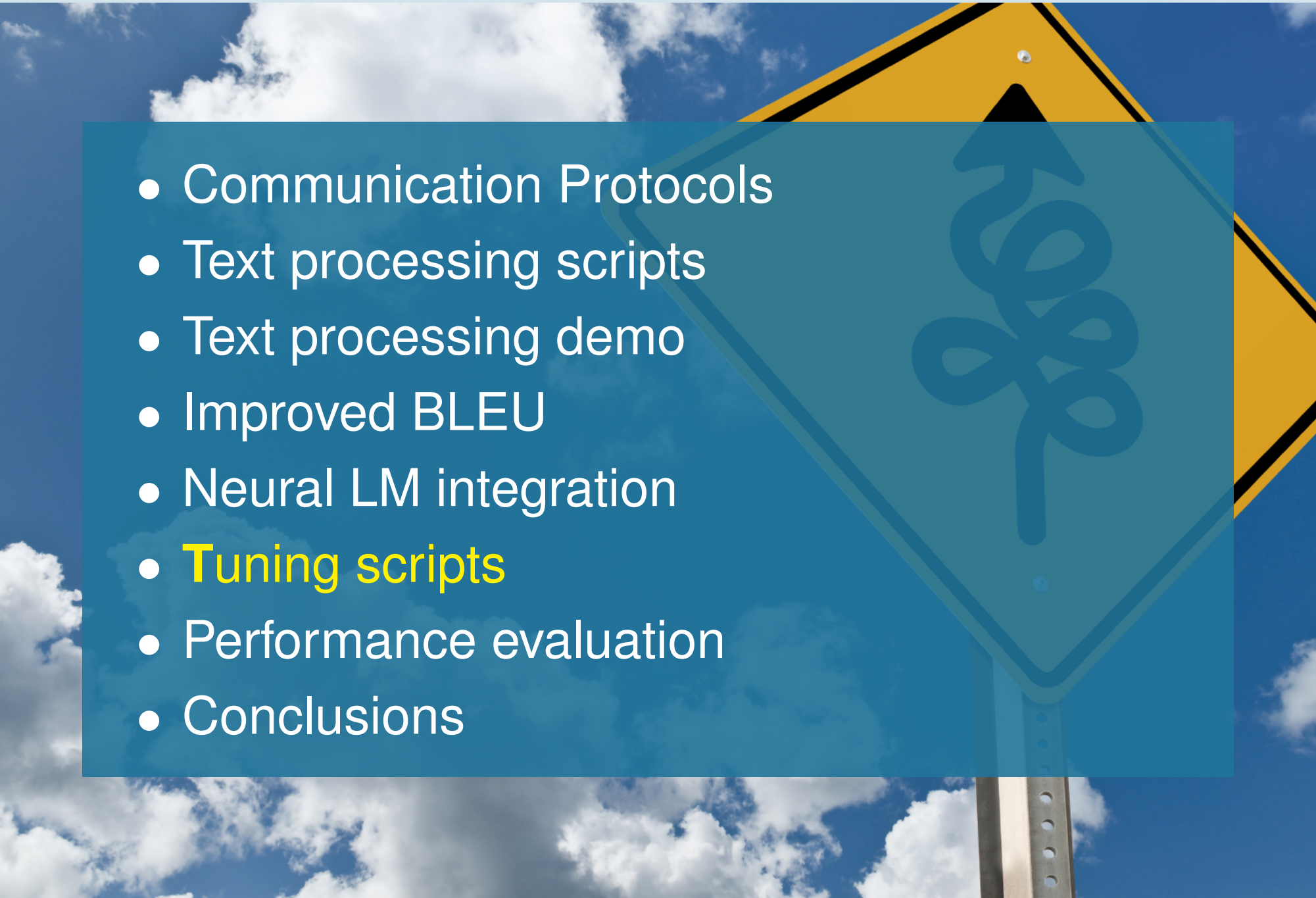


Neural LM integration



Hamidreza Ghader, Neural LM model, just 2 modifications

Table of Contents

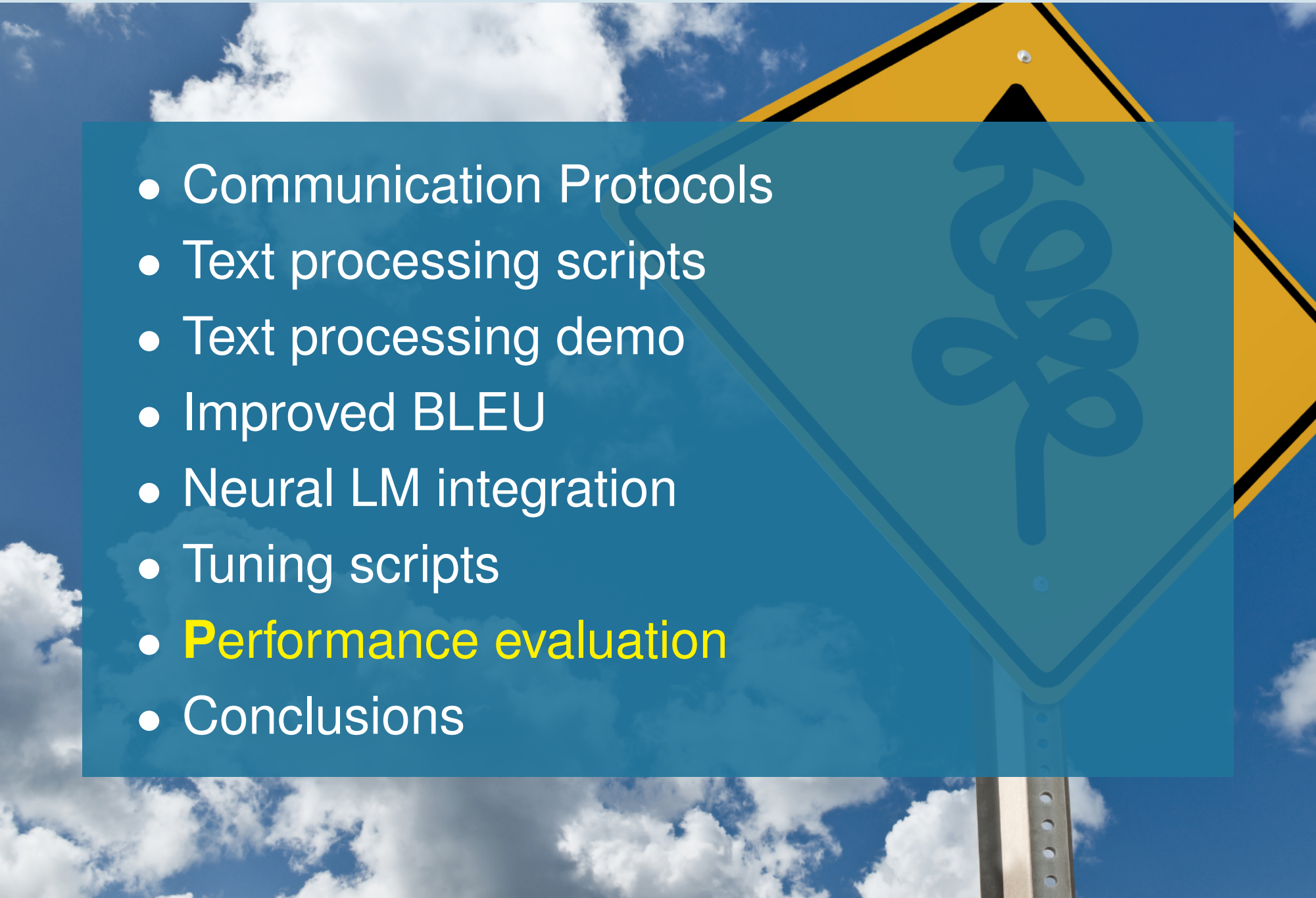


Available Tuning scripts

Tuning: Employ Discriminative Training to optimize translation performance

- `run_tuning.sh`
  - Run tuning infrastructure
- `tuning_progress.pl`
  - Monitor BLEU scores
  - Extract configuration files
- `kill_tuning.pl`
  - Kill tuning processes

Table of Contents



Evaluation: Task and Tools

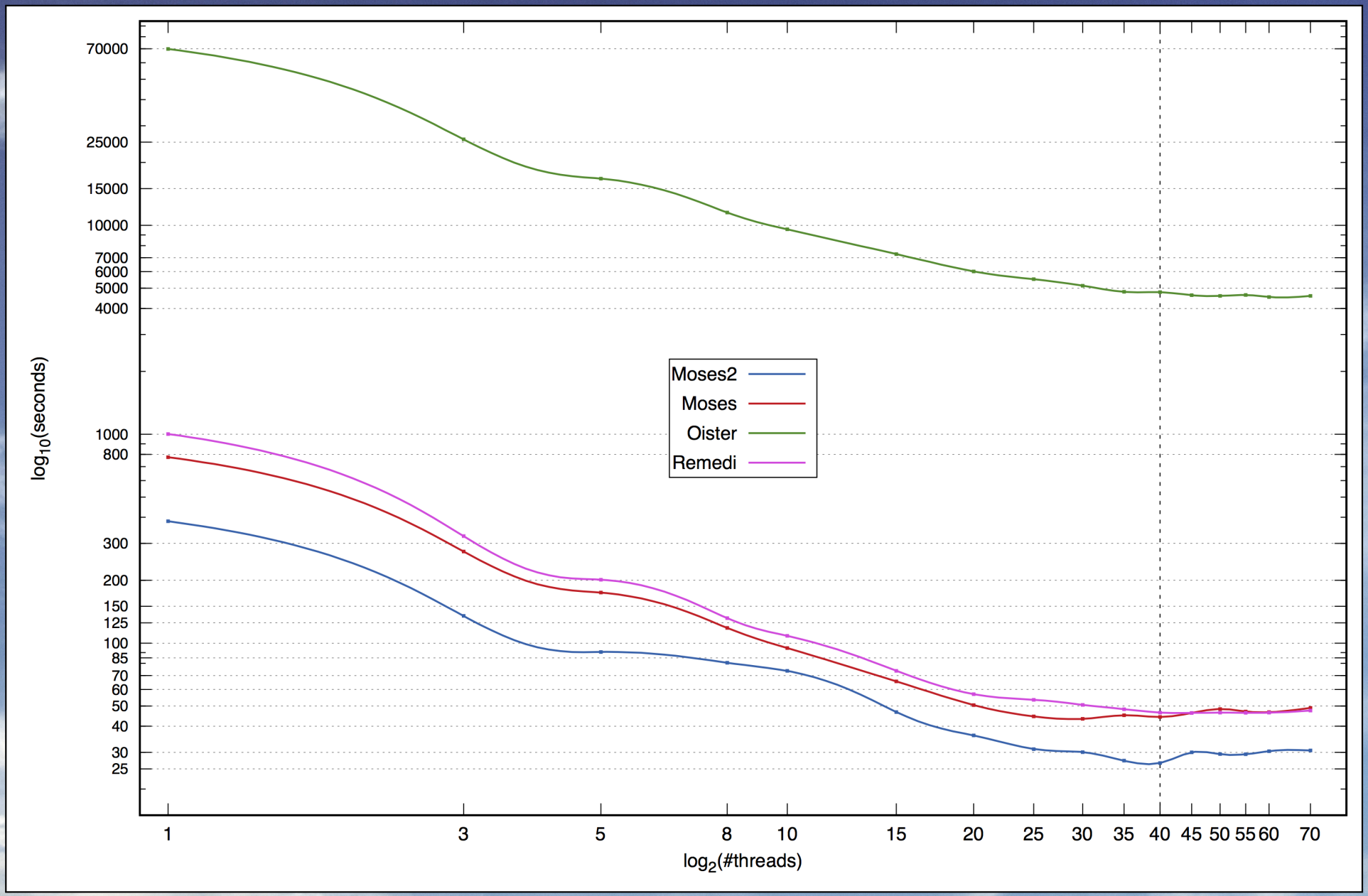
- Multi-threaded SMT systems
  - REMEDI - 2 years old, C++
  - Moses/Moses2 - 11 years old, C++
  - Oister - 6 years old, Perl
- Empirically compare
  - Pure decoding times
  - Scaling on multi-core

Evaluation: Experimental setup

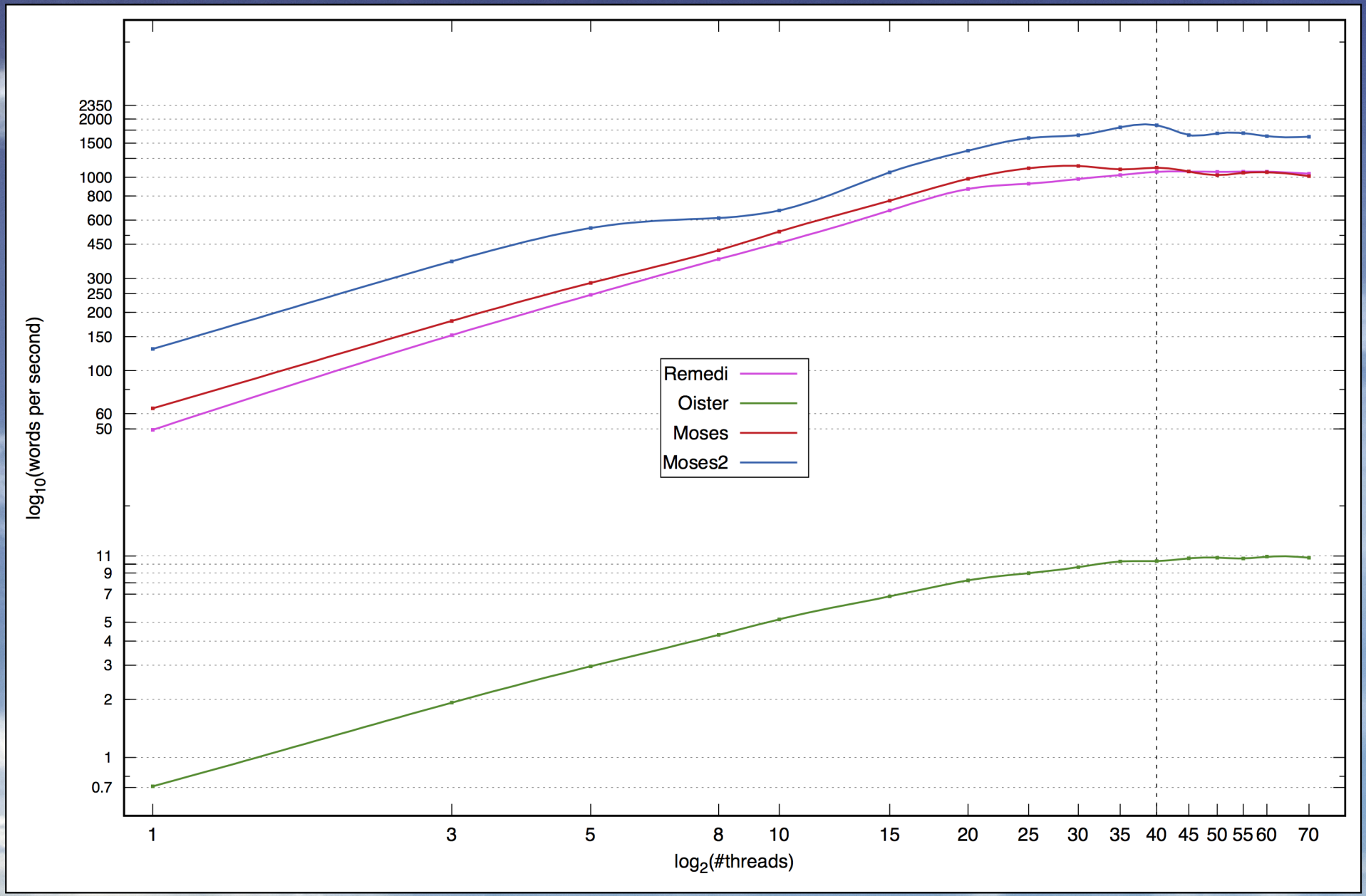
- Chinese to English:
  - MT-04 data set
  - CTB segmented
  - 1788 sent
- Independently tuned:
  - Oister: 36.8 BLEU
  - REMEDI: 36.72 BLEU
  - Moses/2: 35.53 BLEU
- Perform:
  - 10 runs per exp.
  - Oister 3 runs

- Large Models:
  - LM: 48.9 Gb, 5g
  - RM: 9.7 Gb, 8f
  - TM: 1.3 Gb, 5f
- Black-box experiments
- Measure runtime
- Obtain data:
  - Avg loading times
  - Avg total times
  - Std. Deviations

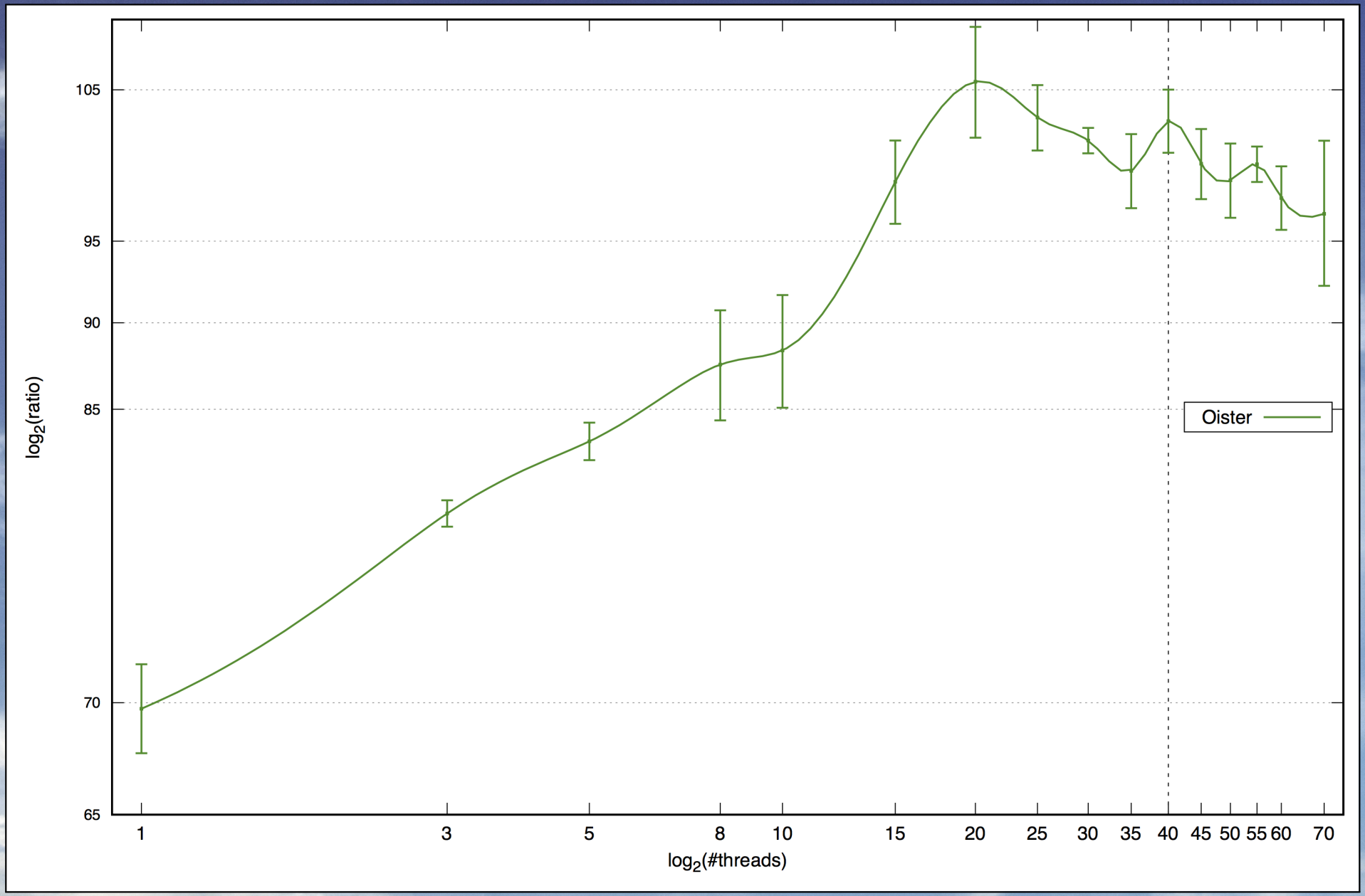
Evaluation: Decoding times



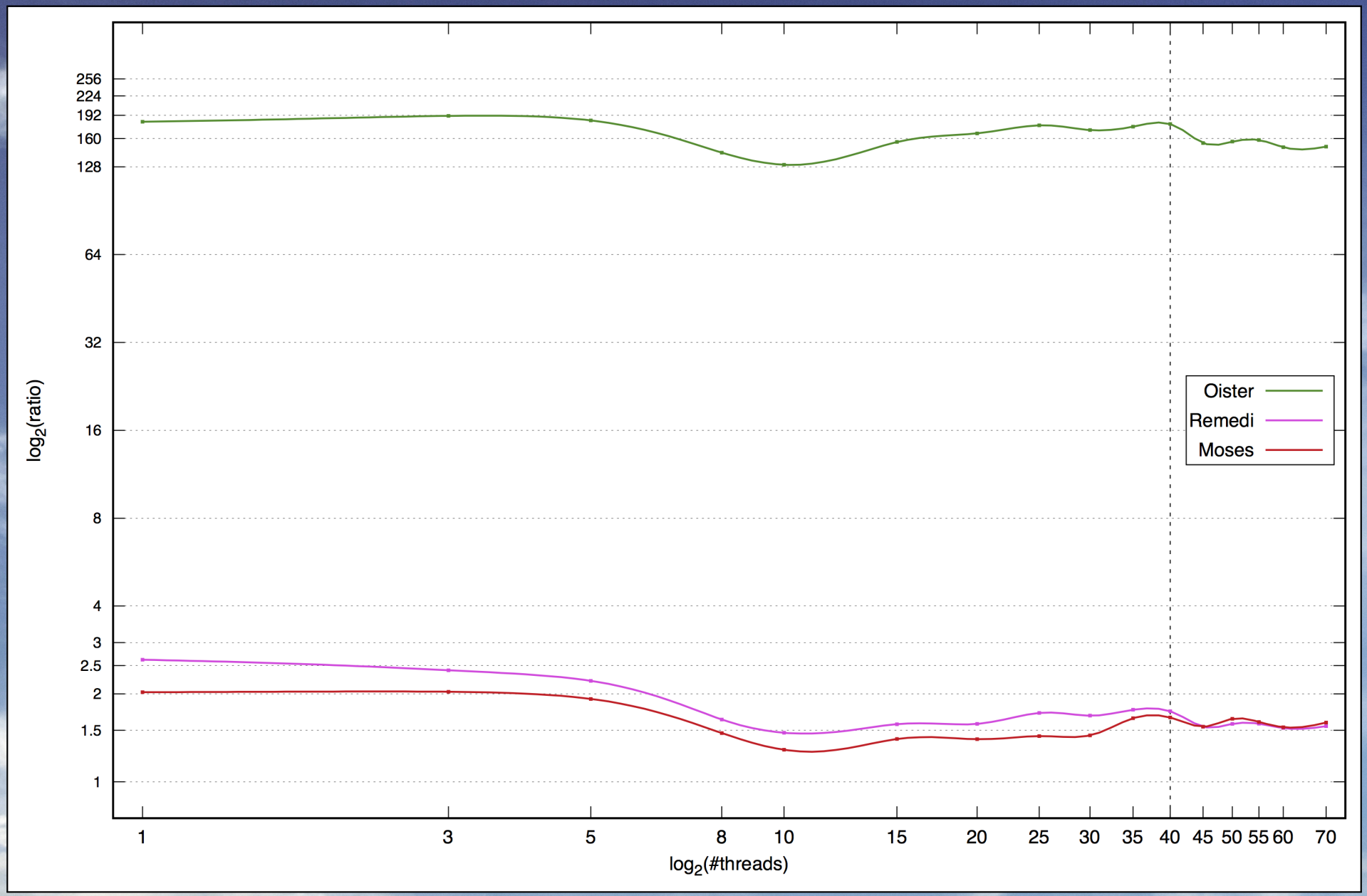
Evaluation: Words per second



Evaluation: Oister vs. Remedi



Evaluation: All vs. Moses2



Evaluation: Multi-threading speedups

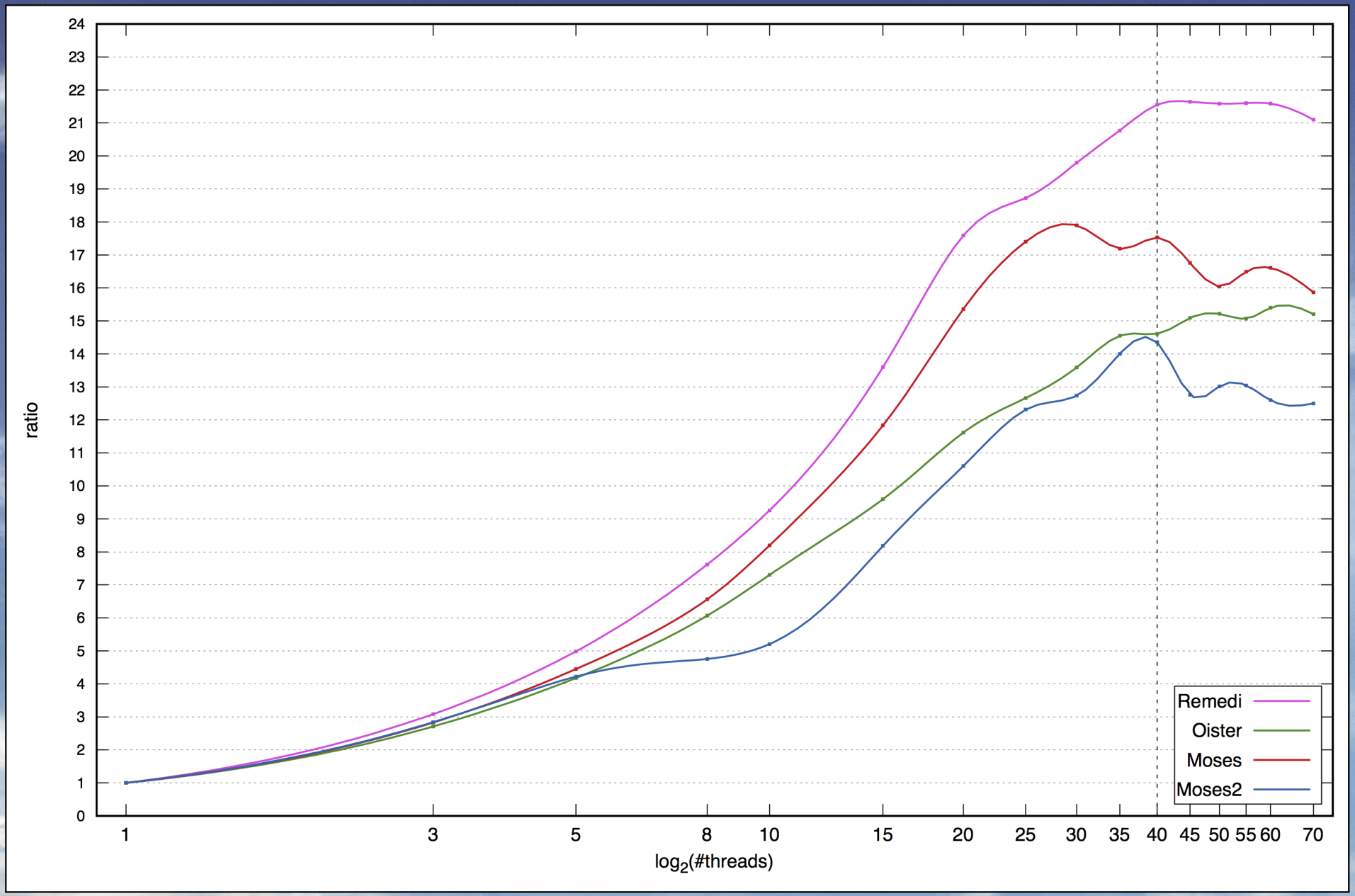
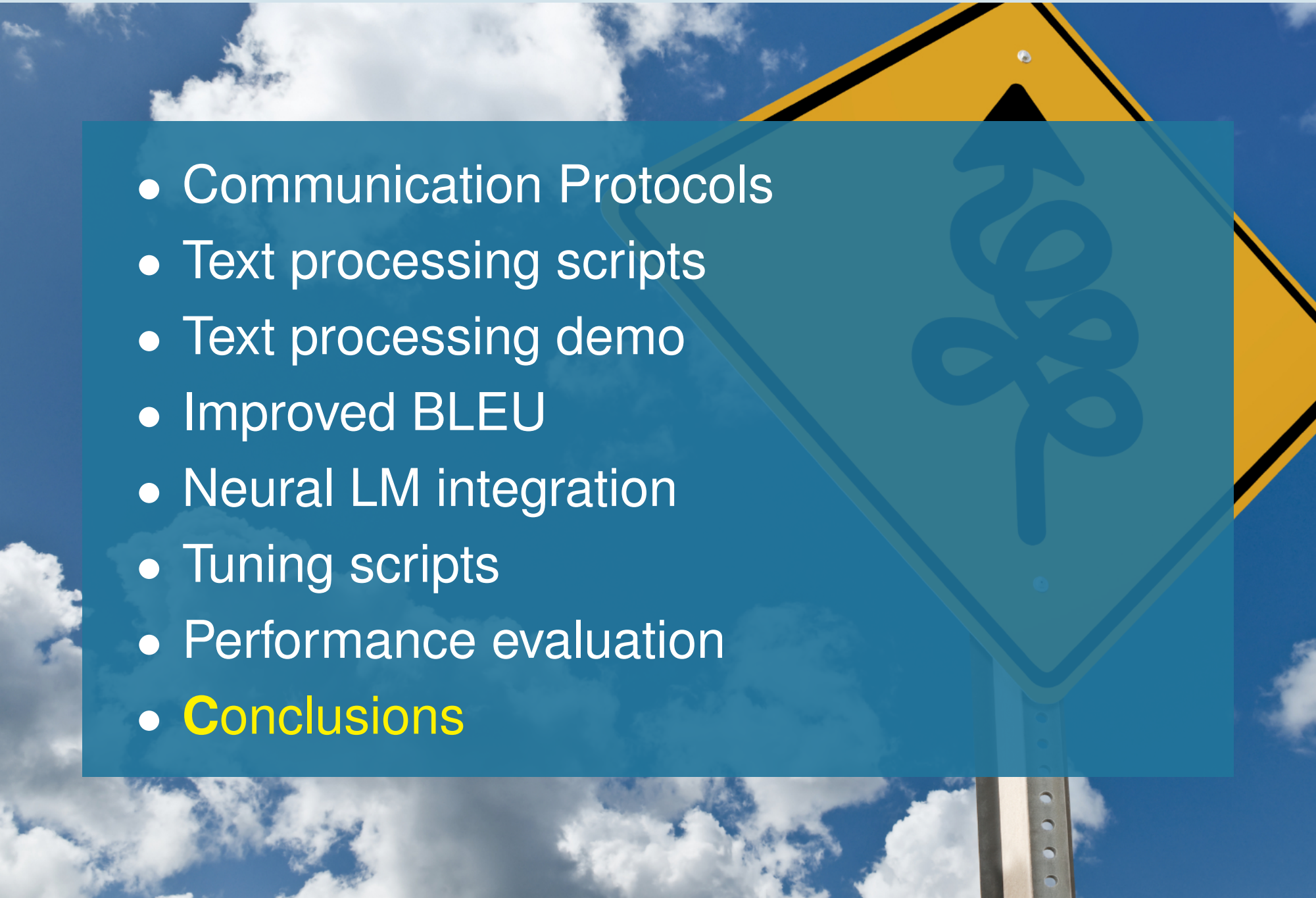
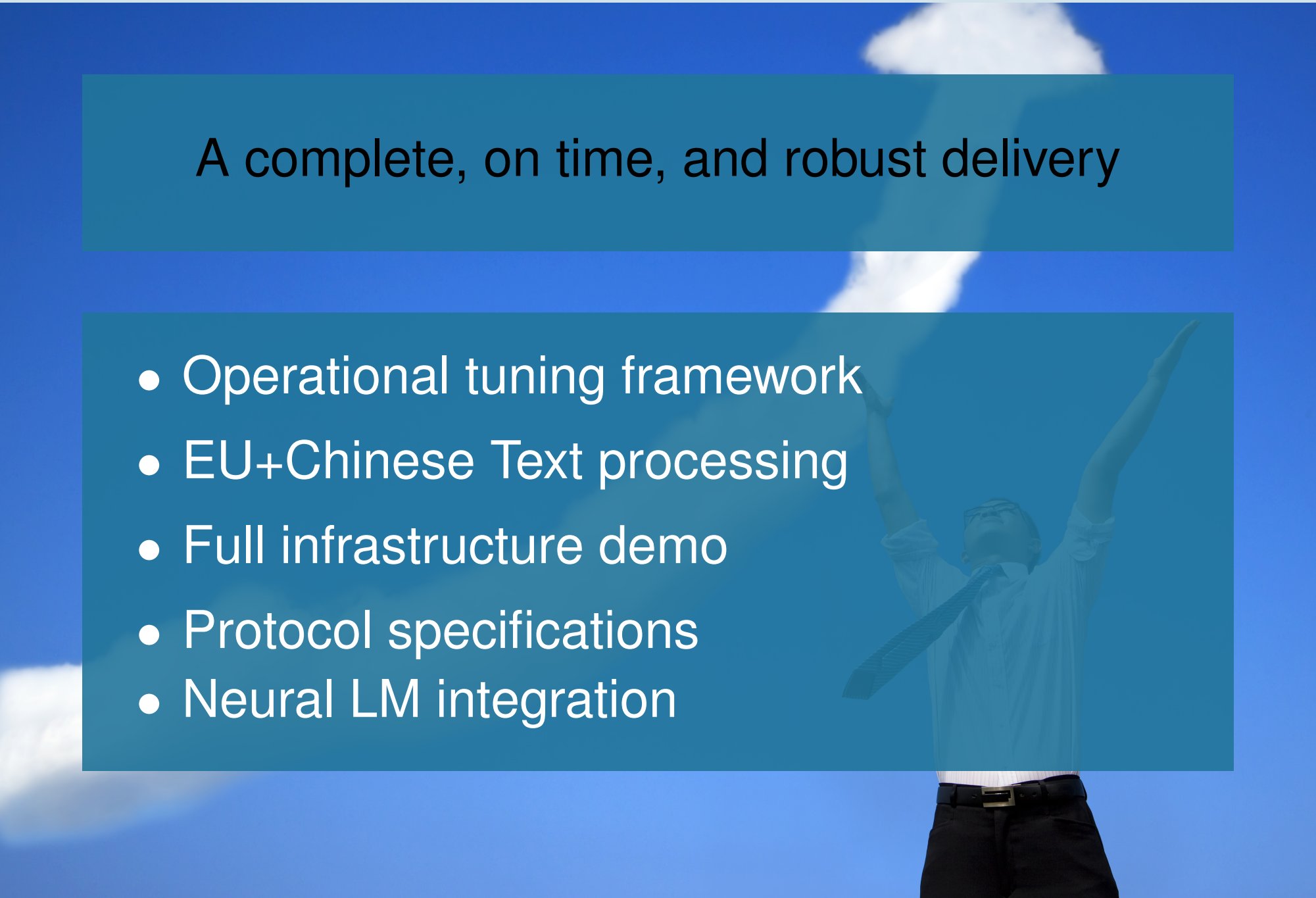


Table of Contents



Conclusions - 1/2



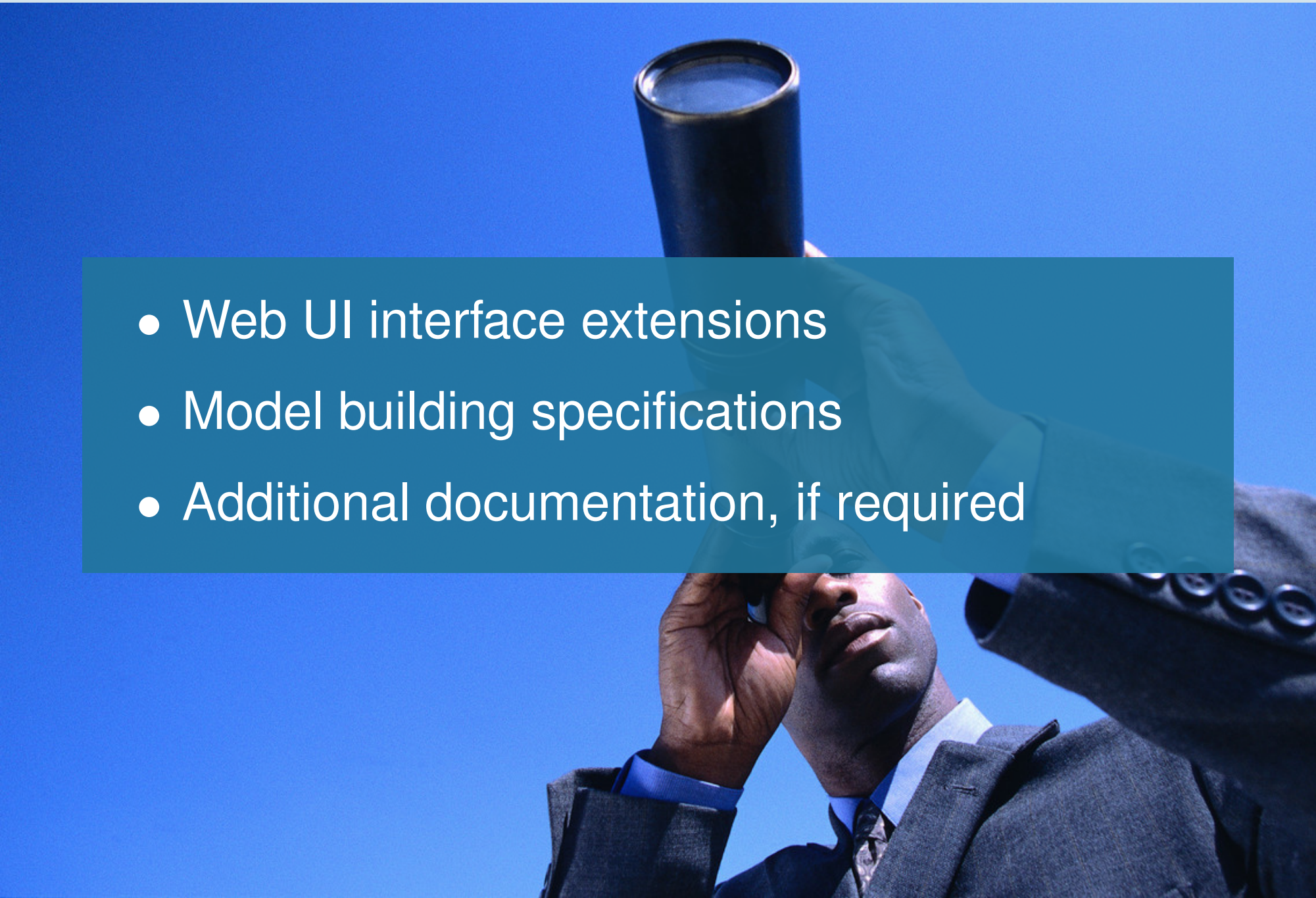
A complete, on time, and robust delivery

- Operational tuning framework
- EU+Chinese Text processing
- Full infrastructure demo
- Protocol specifications
- Neural LM integration

Conclusions - 2/2

- MT-04, Chinese $\rightarrow$ English
  - From $\approx$ 23 $\rightarrow$ 36.72 BLEU
  - Matching Oister in BLEU
- The best scaling in #threads
- Most stable in its decoding times
- 70 to 100 times faster than Oister
- Matches Moses if #threads == #cores
- Is just $\approx$ 2.5 to 1.5 times slower than Moses2
- Better BLEU than Moses/Moses2

Future plans



- Web UI interface extensions
- Model building specifications
- Additional documentation, if required

Discussion

Thank you all!

Literature I

Tamchyna Aleš, Dušek Ondřej, Rudolf Rosa, and Pavel Pecina.
Mtmonkey: A scalable infrastructure for a machine translation web service.
The Prague Bulletin of Mathematical Linguistics, 100:31–40, October 2013.
https://github.com/ufal/mtmonkey.

Truica Ciprian-Octavian, Velcin Julien, and Boicea Alexandru.
Automatic language identification for romance languages using stop words and diacritics.
2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 00:243–246, 2015.

NIST Multimodal Information Group.
Nist 2004 open machine translation (openmt) evaluation, 2010.
http://www.itl.nist.gov/iad/mig/tests/mt/2004/.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst.
Moses: Open source toolkit for statistical machine translation.
In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Edward Loper and Steven Bird.
Nltk: The natural language toolkit.
In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla.
tRuEcasing.
In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 152–159, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Literature II

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky.

The Stanford CoreNLP natural language processing toolkit.
In Association for Computational Linguistics (ACL) System Demonstrations, pages 55–60, 2014.