MultiplEYE Final Conference
August 20-21, 2026
Skopje, North Macedonia

Surprisal by Design: A Principled Pipeline for Human-LLM Stimulus Pairs in Eye-Tracking Reading Experiments

Isabelle C. R. Cretton[*1], **David R. Reich**[1,2], **Lena A. Jäger**[1]

[1]Department of Computational Linguistics, University of Zurich
[2]Department of Computer Science, University of Potsdam

*contact: isabellecarolinerose.cretton@uzh.ch

Behavioural studies of reading often rely on surprisal as a linking hypothesis between language model predictions and fixation durations (Smith and Levy, 2013; Levy, 2008). This framework was originally operationalised using n-gram and probabilistic phrase-structure models of human language (Hale, 2001; Boston *et al.*, 2008; Demberg and Keller, 2008). Yet no principled methodology exists for constructing eye-tracking stimuli that isolate generation source as an experimental variable. Existing human vs LLM text corpora are observational: uncontrolled for topic, length, register, and crucially, not guaranteed to exhibit a positive predictability gap. We present a pipeline that treats stimulus construction as an engineering problem with explicit quality constraints, and solves it.

The pipeline is built around several interlocking design principles that distinguish it from ad hoc generation approaches. Surprisal delta, the difference in bits/token between human and generated text, is estimated by two independent judges that play no role in generation, any pair of language models can serve these roles; a hard invariant enforces agreement at the pair level, making the primary experimental variable model-agnostic by construction. The pipeline provides a framework for diagnosing and addressing corpus-specific failure modes through targeted pre- and post-generation filters; two instantiations are demonstrated here. In factual, entity-dense prose, formulaic register systematically inverts the surprisal direction under news-trained language models, caught by a surprisal-based pre-filter before generation. In L2 argumentative writing, the predictability gap is structurally guaranteed but generation tends toward register drift and essay-mode artefacts, caught by a pattern-level validation layer after generation, with filter logic modular and extensible to any prose type. In this instantiation, generation across both corpora is pooled across model scales and the best output per source text is selected by a composite score weighting surprisal strength, length fidelity, semantic overlap, syntactic similarity, and information density uniformity. Scientific constraints are enforced as hard blocking assertions in a dedicated test suite rather than as post-hoc recommendations, and full reproducibility is ensured via a seeded generation chain spanning sampling, model inference, and dataset construction.

As a proof of concept applied to two typologically distinct corpora, the pipeline yields fully validated stimulus pairs with a mean surprisal delta of 0.957 bits/token, dual-judge agreement on every pair, BERTScore F1 of 0.821, and length ratios above 0.87, all passing a blocking test suite that enforces scientific invariants as hard assertions rather than recommendations. The pipeline is corpus-agnostic, fully reproducible via seeded generation, and releases both stimuli and tooling for the community.

References

Boston M. F., Hale J. T., Kliegl R., Patil U., Vasishth S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research 2*. doi:https://doi.org/10.16910/JEMR.2.1.1.

Demberg V., Keller F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition 109*(2), 193–210. doi:https://doi.org/10.1016/j.cognition.2008.07.008.

Hale J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, North American Chapter of the Association for Computational Linguistics, USA, pp. 1–8. Association for Computational Linguistics. doi:https://doi.org/10.3115/1073336.1073357.

Levy R. (2008). Expectation-based syntactic comprehension. *Cognition 106*(3), 1126–1177. doi:https://doi.org/10.1016/j.cognition.2007.05.006.

Smith N. J., Levy R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition 128*(3), 302–319. doi:https://doi.org/10.1016/j.cognition.2013.02.013.