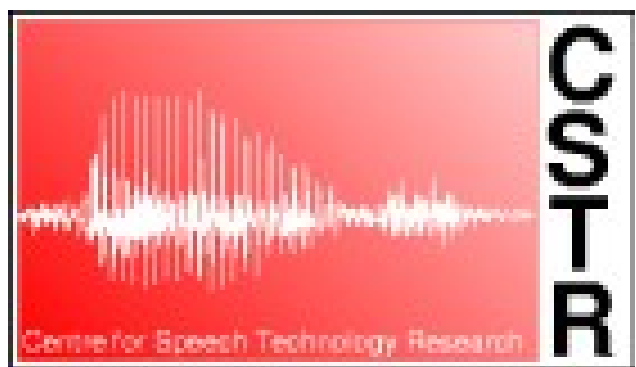
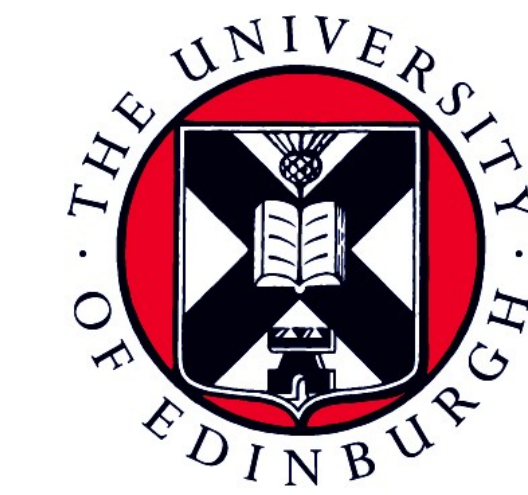
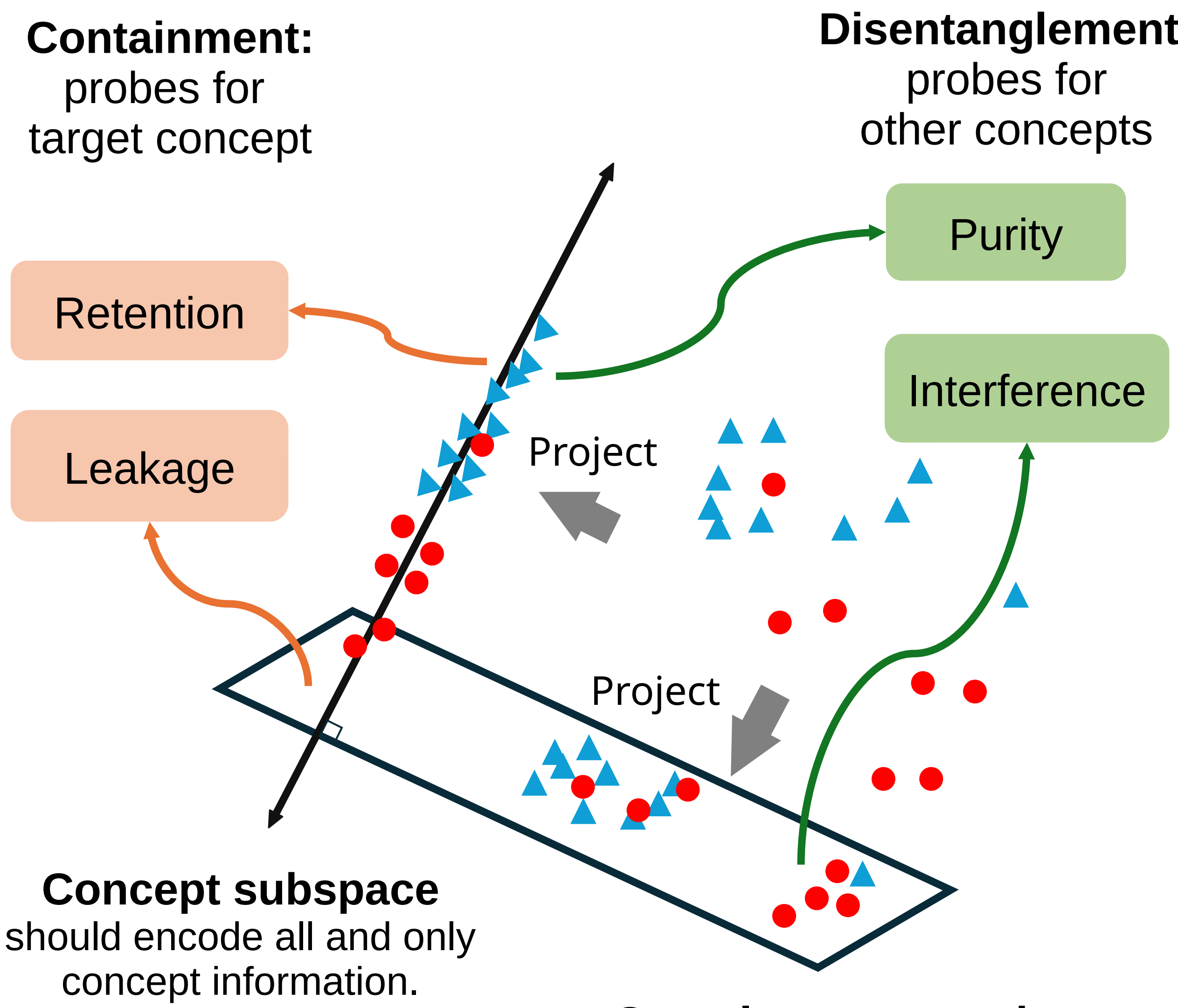
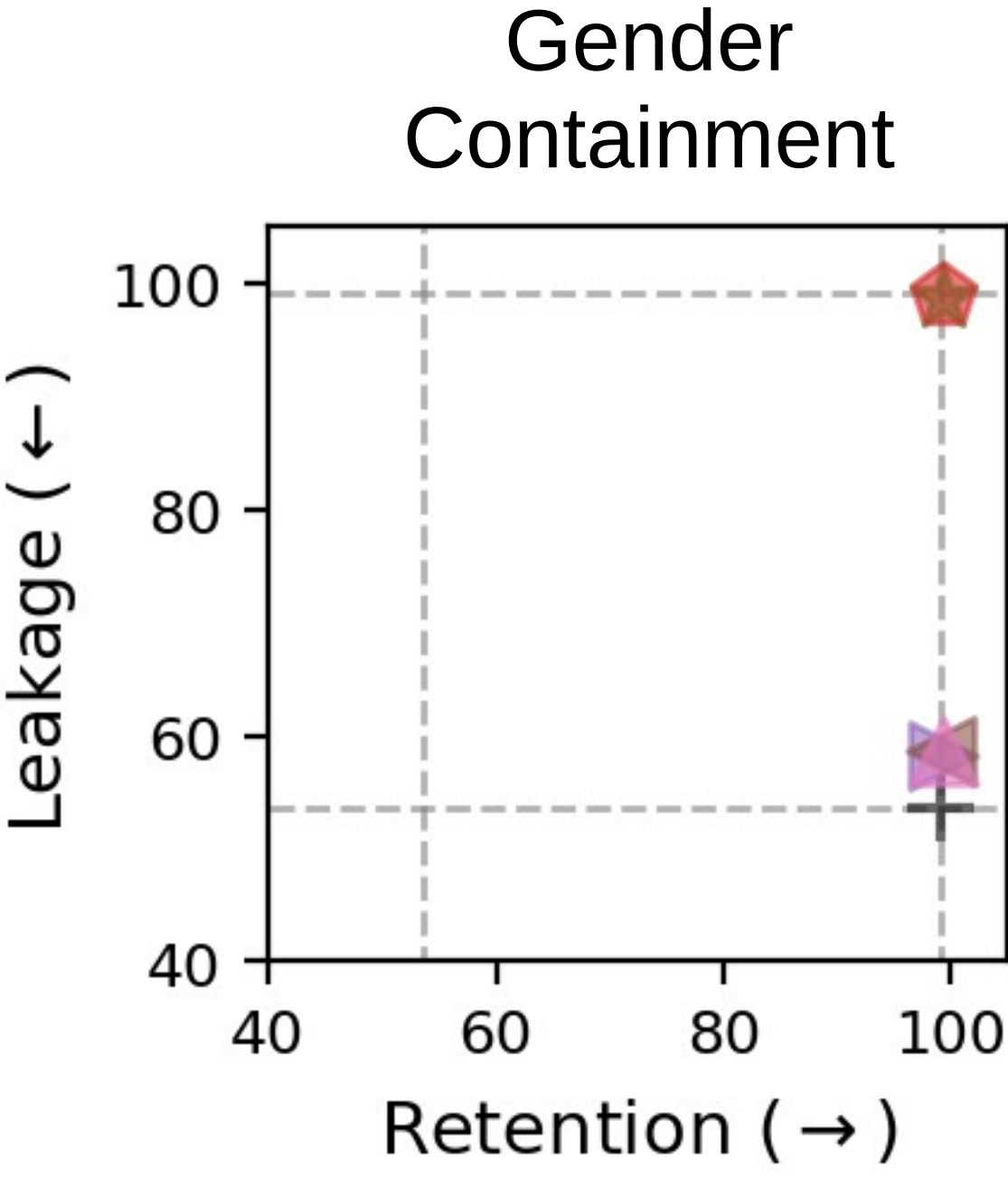
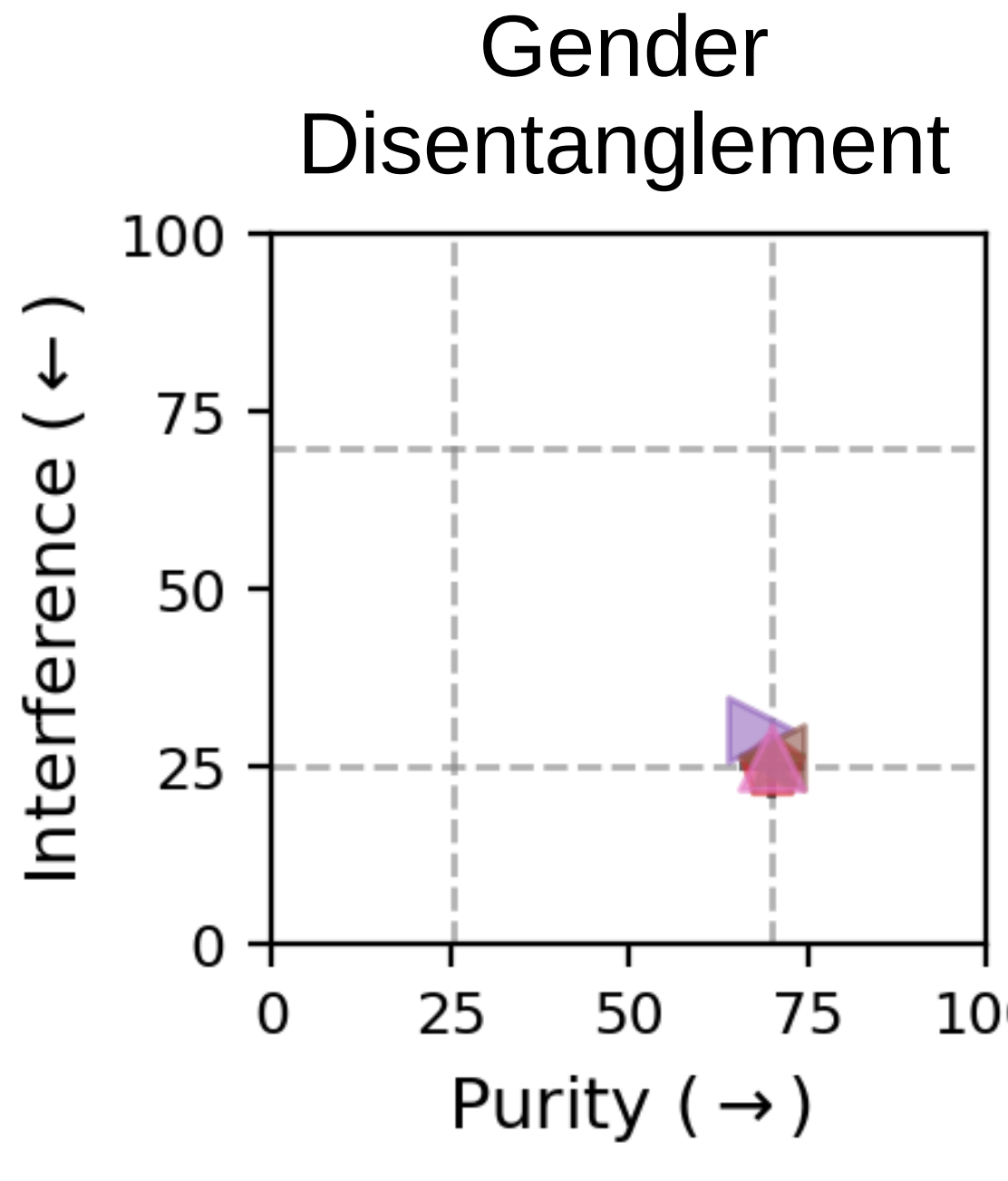
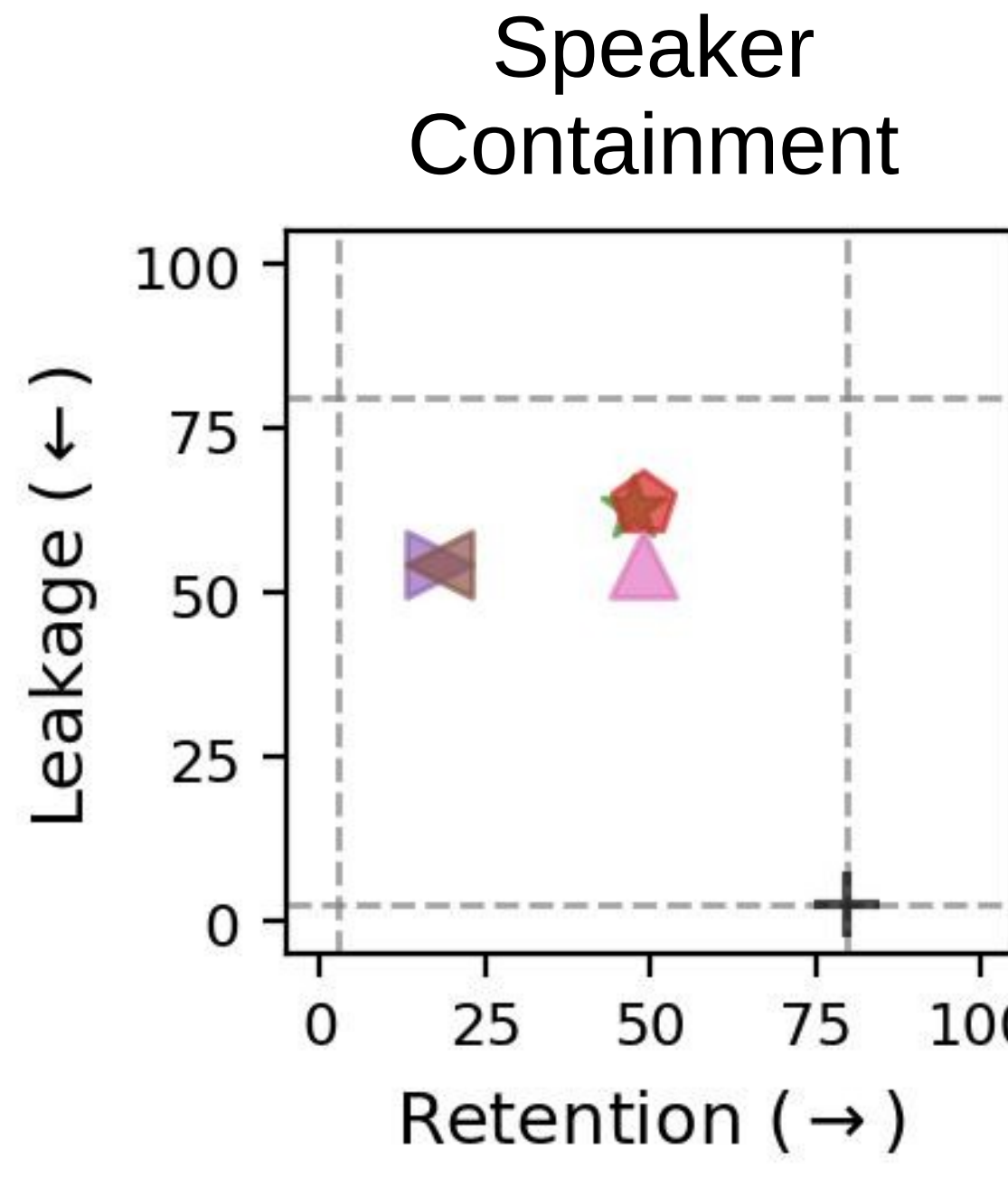
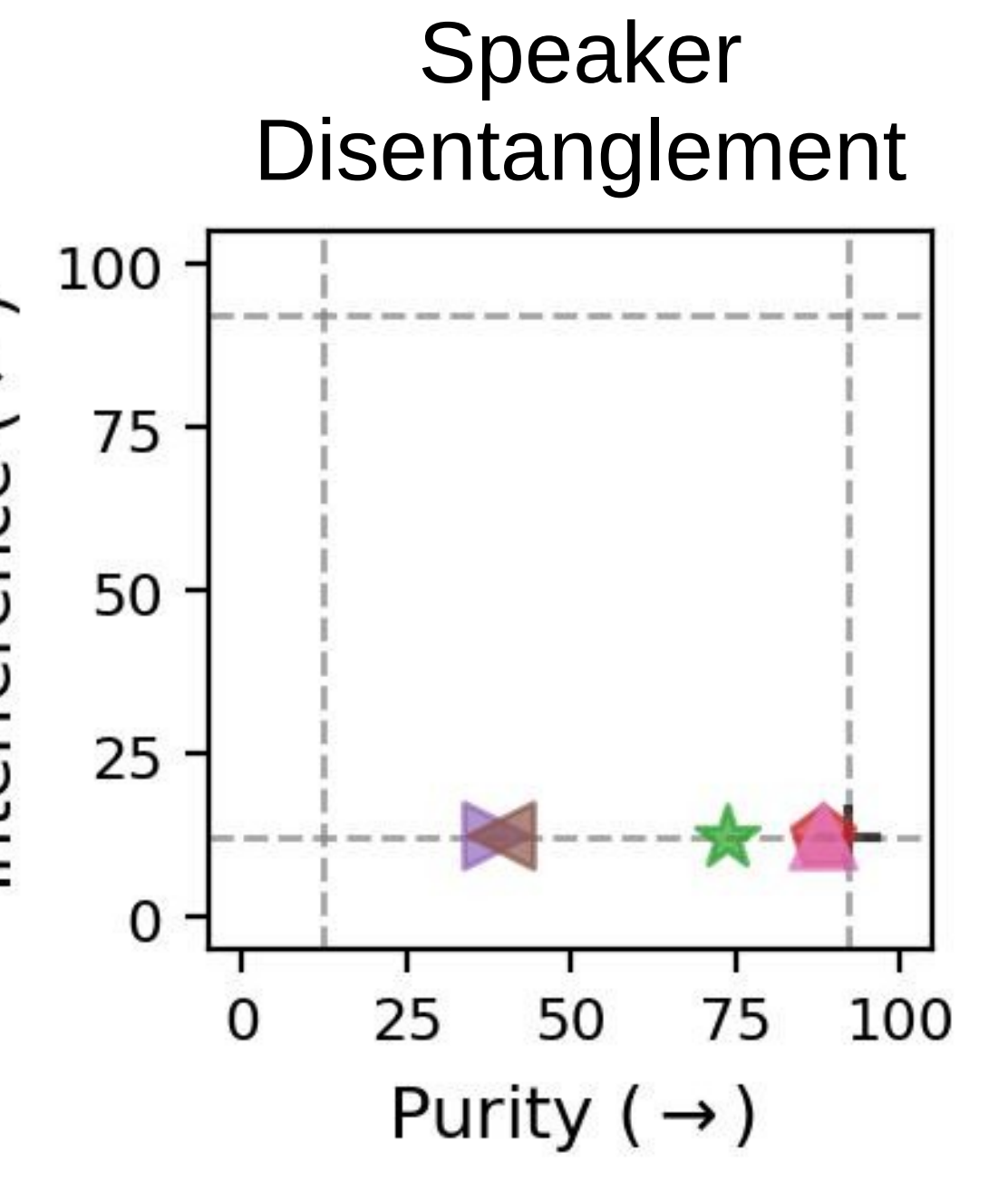


# A framework for analyzing concept representations in neural models



Burin Naowarat, Hao Tang, Sharon Goldwater  
Centre for Speech Technology Research, School of Informatics, University of Edinburgh

| I) Findings                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | III) The proposed framework                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>• Concept subspaces depend on the estimation method.</li><li>• So may structural interpretations.</li></ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <p><b>We proposed two concept subspace measures:</b></p> <ul style="list-style-type: none"><li>• <b>Containment.</b> How well is the target concept encoded in the concept subspace, and not outside it?</li><li>• <b>Disentanglement.</b> How well is a concept subspace disentangled from other concepts?</li></ul> <p>We quantify them using four (linear) probes, defined over a linear decomposition of the representation space.</p>  <p><b>Containment:</b> probes for target concept</p> <p><b>Disentanglement:</b> probes for other concepts</p> <p><b>Retention</b></p> <p><b>Leakage</b></p> <p><b>Project</b></p> <p><b>Concept subspace</b> should encode all and only concept information.</p> <p><b>Complementary subspace</b> should preserve all information about other concepts.</p> <p><b>Purity</b></p> <p><b>Interference</b></p>                                                                                                             |
| II) Motivation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | IV) Different estimators find different concept subspaces.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <p><b>Why study concept representations in models?</b></p> <ul style="list-style-type: none"><li>• Understand information organization (Chang et al., 2022)</li><li>• A proxy for human brains (Liu et al., 2024, Zhang et al., 2026)</li></ul> <p><b>Previous work uses concept subspaces to interpret concept representations.</b></p> <ul style="list-style-type: none"><li>• Estimate <i>one</i> concept subspace, and analyze its structures.</li></ul> <p><b>Example findings:</b></p> <ul style="list-style-type: none"><li>• Token position information was encoded along toruses, spirals, and curves (Chang et al., 2022)</li><li>• Speakers and phonetics are encoded in orthogonal dimensions (Liu et al., 2023)</li></ul> <p><b>Concept subspaces can be estimated in different ways. Do they have similar properties?</b></p> <div><div> MLR<br/> LDA</div><div> CPCA<br/> COV<br/></div></div> <p>Probing subspaces (Chang et al., 2022)</p> <p>Erasure subspaces (Liu et al., 2023, Belrose et al., 2023)</p> | <p><b>Language model:</b> <i>gender</i> vs. <i>profession</i> concepts.<br/>Gender erasure subspaces have better gender containment.</p> <div><div><p>Gender Containment</p></div><div><p>Gender Disentanglement</p></div></div> <p>Model input: A short paragraph of biography (Bias in bios)<br/>Representation: BERT last-layer CLS tokens (sequence-level)</p> <p><b>Speech model:</b> <i>phonetic</i> vs. <i>speaker</i> concepts.<br/>LEACE speaker subspace have the best overall property.</p> <div><div><p>Speaker Containment</p></div><div><p>Speaker Disentanglement</p></div></div> <p>Model input: Read speech (LibriSpeech)<br/>Representation: HuBERT layer-11 outputs (frame-level)</p> |
| <p>• Liu, Oli Danyi, et al. "A predictive learning model can simulate temporal dynamics and context effects found in neural representations of continuous speech." <i>Cogsci</i> (2024).</p> <p>• Zhang, Yizhen, et al. "Human Cortical Dynamics of Auditory Word Form Encoding." <i>Neuron</i> (2026).</p> <p>• Chang, Tyler, et al. "The geometry of multilingual language model representations." <i>EMNLP</i> (2022).</p> <p>• Liu, Oli, Hao Tang, and Sharon Goldwater. "Self-Supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces." In <i>Interspeech</i> (2023).</p> <p>• Belrose, Nora, et al. "LEACE: Perfect Linear Concept Erasure in Closed Form." <i>NeurIPS</i> (2023).</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |