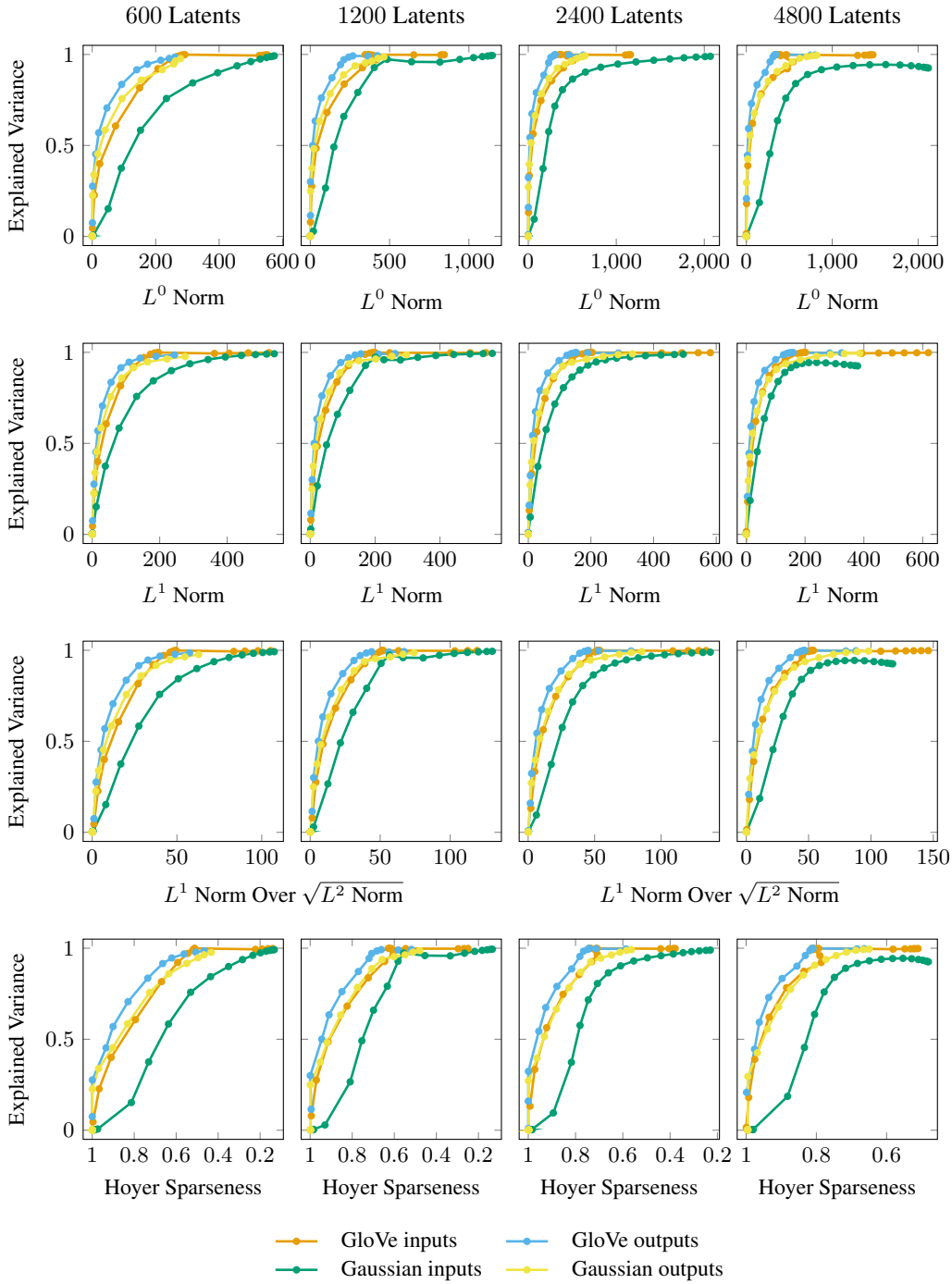


Figure 29: Pareto frontiers of explained variance against sparsity measures for 300-dimensional GloVe word embeddings, Gaussian controls with the same mean and variance, and the corresponding outputs when these are passed to a randomly initialized two-layer MLP.

## J EXAMPLE FEATURES FOR PYTHIA-6.9B

VARIANT: TRAINED

FEATURE 180935 (LAYER 0)

**Interpretation:** The term "security" is predominantly used to refer to protection, safety, and measures to prevent harm, while "oz" is likely referring to ounces, possibly in a context of measurement or quantification, although "oz" appears less frequently and often in a different context.

**Top Examples:**

1. Text: <endoftext—¿—. If for any reason you are unhappy with our service please contact us directly so we can make it right for you.Journal of Cyber Security, Vol.
   Activation: 4.3750
   Active tokens: Security
2. Text: <endoftext—¿— Security Practitioner. Tremendously passing CompTIA Advanced Security Practitioner (casp) cert has never been as easy as it
   Activation: 4.3125
   Active tokens: Security Security
3. Text: in trusted hands for your Cyber Security career or staffing needs. Call 0203 643 0248 to find out more. Technically proficient using
   Activation: 4.3125
   Active tokens: Security

---

FEATURE 93790 (LAYER 8)

**Interpretation:** Nouns and phrases related to economic concepts, development, and business, often referring to growth, progress, and improvement.

**Top Examples:**

1. Text: training requirements. See "Workforce" section for additional information. The Economic Development Transportation Fund, commonly referred to as the "Road Fund," is an
   Activation: 21.3750
   Active tokens: Development
2. Text: Montréal.The Williamsburg Economic Development Authority offers a 33% matching grant up to $7,500 for exterior improvements to existing businesses in the City of
   Activation: 20.2500
   Active tokens: Development Authority
3. Text: Correction: In a July 16 web story The Real Deal incorrectly stated that the Economic Development Corporation was "circumventing" laws with its restructuring. In
   Activation: 20.0000
   Active tokens: Development

---

FEATURE 128309 (LAYER 12)

**Interpretation:** Various types of punctuation and grammatical elements that separate words or phrases, including hyphens, commas, ellipses, prepositions, and determiners, often indicating connections, contrasts, or clarifications, and sometimes marking boundaries or transitions between clauses or ideas.

**Top Examples:**

1. Text: Run it in JDK6, and it will print "[axons, bandrils, chumblies]". If you are having trouble switching from
   Activation: 8.0000
   Active tokens: in JD K

2. Text: Here, we introduce the coordinate systems for three-dimensional space □□□2. The study of 3-dimensional spaces lead us to the setting for our study
Activation: 7.8125
Active tokens: □

3. Text: .path.expanduser("˜/malwarehouse/") because this server doesn't have X-Windows running. If you are looking for a simple and
Activation: 7.7188
Active tokens: . path expand user

---

VARIANT: STEP 0

FEATURE 126848 (LAYER 12)

**Interpretation:** Nouns denoting people who train others, units or marks of measurement, and abbreviations or acronyms representing specific standards or technologies.

**Top Examples:**

1. Text: What are the various lessons a member can access at a tennis club? Whether you are a beginner or advanced player, trainers help you to choose the right gaming
Activation: 13.1250
Active tokens: trainers

2. Text: report include various simulation platforms and Serious Games. The report also analyzes some major allied products such as patient simulators and task trainers. The technologies analyzed
Activation: 13.0625
Active tokens: trainers

3. Text: a stylish spring in your step when you buy from our fantastic range of men's and women's Asics trainers. We've got numerous styles from
Activation: 13.0000
Active tokens: trainers

---

FEATURE 2125 (LAYER 4)

**Interpretation:** The word "papers" is often used in contexts referring to written documents, such as academic papers, court documents, or printed materials, and is frequently mentioned in relation to tasks like writing, research, and education.

**Top Examples:**

1. Text: caustic solution . As an abrasive, alumina is coated into abrasive papers and .. Pakistan. Sierra leone. Taiwan. Turkey. Venezuela.
Activation: 5.7188
Active tokens: papers

2. Text: that can be associated with interaction with other individuals. For everybody who is uncertain regardless of whether your papers is misstep no cost, buy inexpensive experienced proofreading services
Activation: 5.6875
Active tokens: papers

3. Text: who RV, often traveling in groups, often alone. You just want to have all the papers like RC, licence and insurance coverage as effectively as PUC (
Activation: 5.6562
Active tokens: papers

---