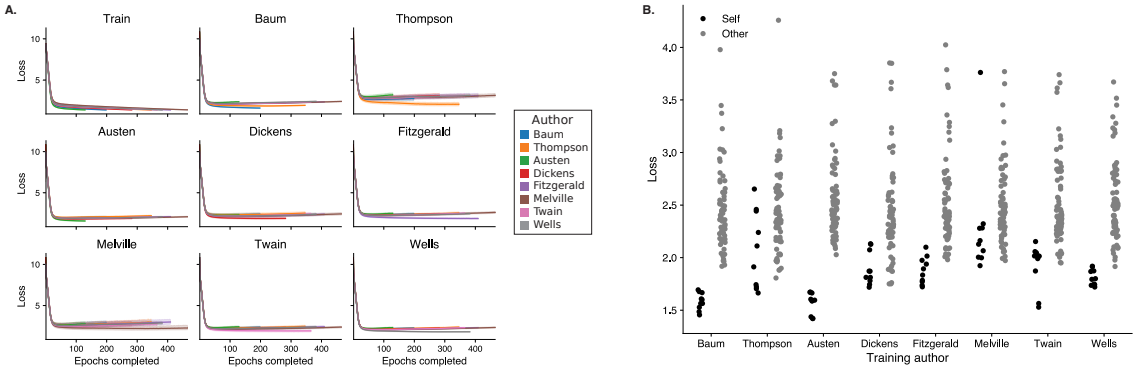
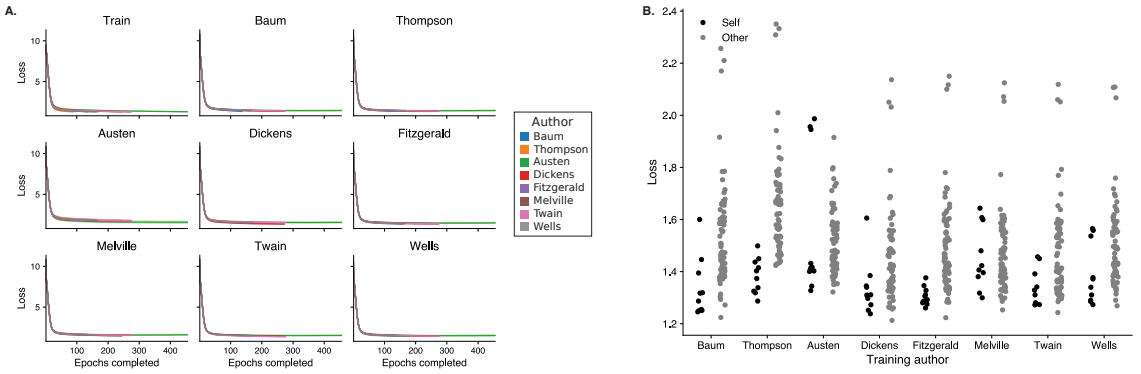
Supplementary materials for: A Stylometric Application of Large Language Models

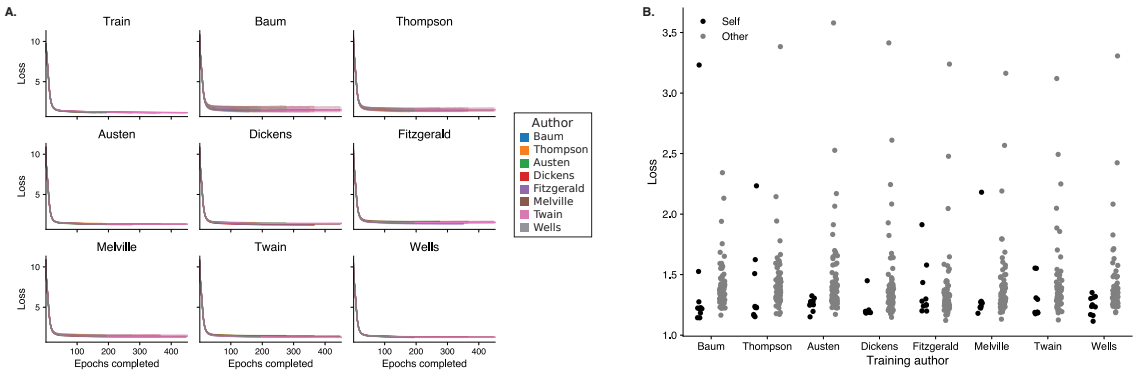
Harrison F. Stropkay, Jiayi Chen, Mohammad J. Latifi,
Daniel N. Rockmore, and Jeremy R. Manning
Dartmouth College
Hanover, NH 03755, USA
{harrison.f.stropkay.25, jiayi.chen.gr, mohammad.javad.latifi.jebelli
daniel.n.rockmore, jeremy.r.manning}@dartmouth.edu



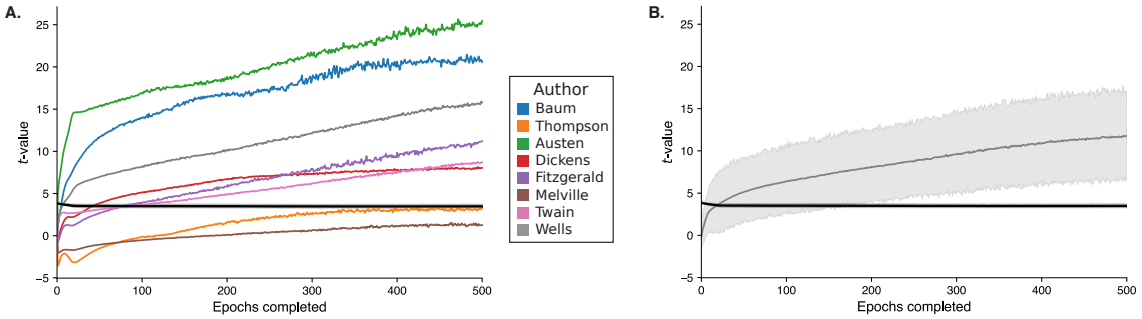
Supplementary Figure 1: Cross-entropy loss across models and authors using only content words. Follows the general format of Figure 1 in the main text, but uses models trained on only content words. All function words are masked out using <FUNC>. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (x-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



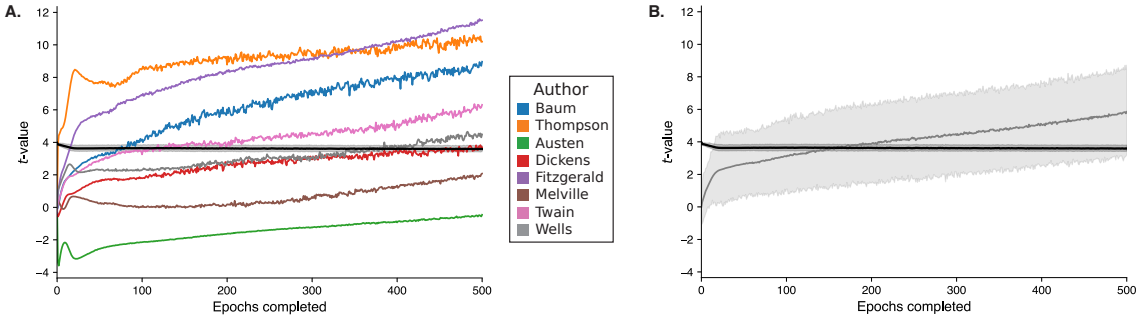
Supplementary Figure 2: Cross-entropy loss across models and authors using only function words. Follows the general format of Figure 1 in the main text, but uses models trained on only function words. All content words are masked out using <CONTENT>. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (*x*-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



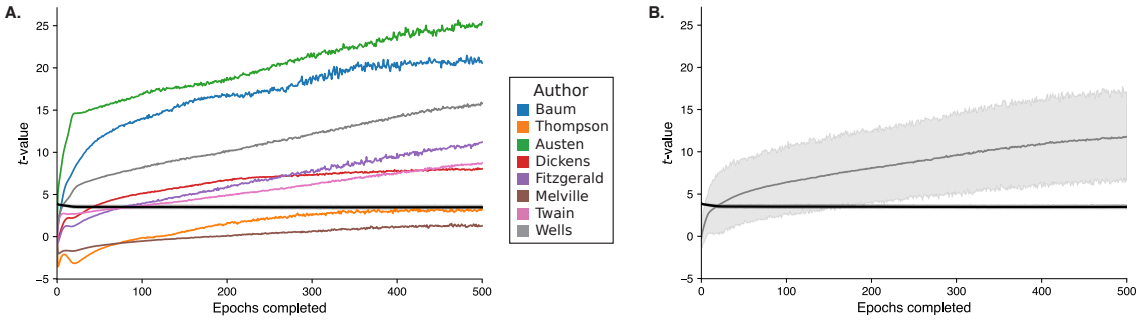
Supplementary Figure 3: Cross-entropy loss across models and authors using only parts of speech. Follows the general format of Figure 1 in the main text, but uses models trained on only parts of speech. All words are replaced with their corresponding part of speech tag. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (*x*-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



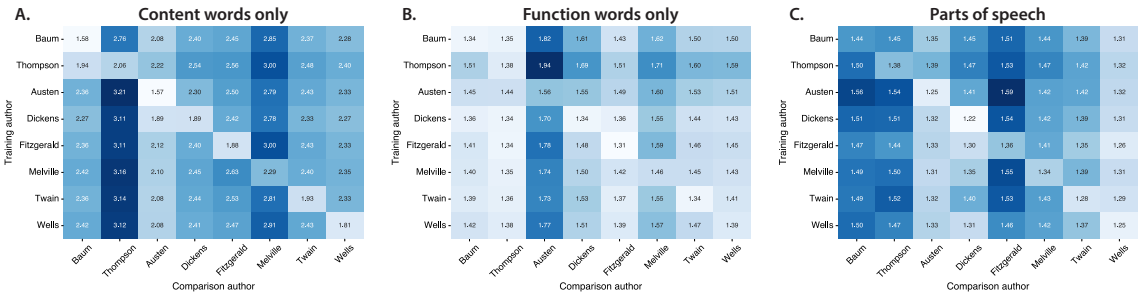
Supplementary Figure 4: Same vs. other author comparisons, by model, using only content words. Follows the general format of Figure 2 in the main text, but uses models trained on only content words. All function words are masked out using <FUNC>. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



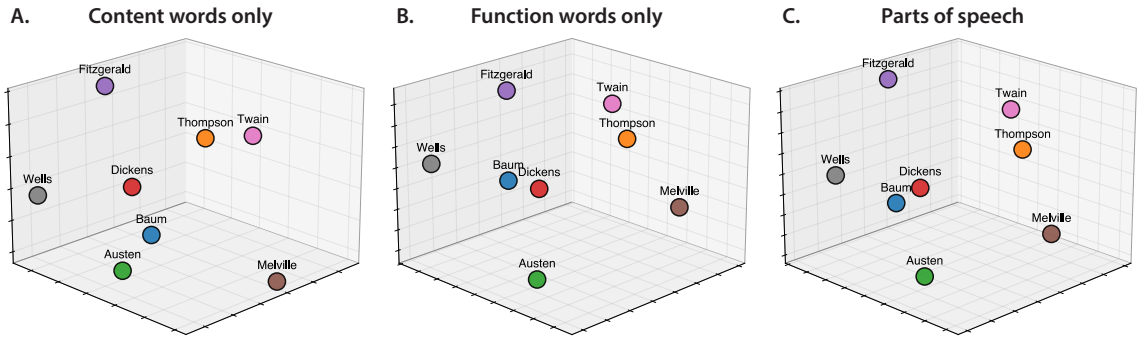
Supplementary Figure 5: Same vs. other author comparisons, by model, using only function words. Follows the general format of Figure 2 in the main text, but uses models trained on only function words. All content words are masked out using <CONTENT>. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



Supplementary Figure 6: Same vs. other author comparisons, by model, using only parts of speech. Follows the general format of Figure 2 in the main text, but uses models trained on only parts of speech. All words are replaced with their corresponding part of speech tag. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



Supplementary Figure 7: Confusion matrices. Follows the general format of Figure 3 in the main text, but shows confusion matrices for models trained on only content words (A), only function words (B), and only parts of speech (C). Within each panel, the matrix displays the average cross-entropy loss assigned by models trained on each author's writing (column) to held-out texts from each author (row), after subtracting the native author's baseline loss.



Supplementary Figure 8: Multidimensional scaling plots. Follows the general format of Figure 4 in the main text, but shows MDS projections of the (symmetrized) average cross entropy loss matrices shown in Figure 7, for models trained on only content words (A), only function words (B), and only parts of speech (C).

Model	t-stat	df	p-value
Baum	20.58	68.36	5.27×10^{-31}
Thompson	3.29	11.33	6.97×10^{-3}
Austen	25.46	70.33	3.20×10^{-37}
Dickens	8.04	37.39	1.13×10^{-9}
Fitzgerald	11.21	49.02	3.97×10^{-15}
Melville	1.28	10.28	0.2274
Twain	8.73	22.50	1.12×10^{-8}
Wells	15.79	71.87	4.53×10^{-25}

Supplementary Table 1: Loss differences between same-author and other-author texts using only content words. Follows the general format of Table 1 in the main text, but uses models trained on only content words. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.

Model	t-stat	df	p-value
Baum	8.97	20.85	1.34×10^{-8}
Thompson	10.20	22.96	5.39×10^{-10}
Austen	-0.46	9.52	0.6581
Dickens	3.69	17.46	1.73×10^{-3}
Fitzgerald	11.52	77.98	1.70×10^{-18}
Melville	2.08	17.29	0.0529
Twain	6.31	34.66	3.14×10^{-7}
Wells	4.49	15.94	3.76×10^{-4}

Supplementary Table 2: Loss differences between same-author and other-author texts using only function words. Follows the general format of Table 1 in the main text, but uses models trained on only function words. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.

Model	t-stat	df	p-value
Baum	0.04	9.22	0.9695
Thompson	1.05	10.91	0.3179
Austen	5.72	77.97	1.89×10^{-7}
Dickens	4.41	62.85	4.14×10^{-5}
Fitzgerald	0.43	14.25	0.6704
Melville	0.81	11.98	0.4337
Twain	2.43	20.88	0.0240
Wells	4.05	56.90	1.55×10^{-4}

Supplementary Table 3: Loss differences between same-author and other-author texts using only parts of speech. Follows the general format of Table 1 in the main text, but uses models trained on only parts of speech. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.