# Toward Compositional Behavior in Neural Models: A Survey of Current Views

**Kate McCurdy**
Universität des Saarlandes
kmccurdy@lst.uni-saarland.de

**Paul Soulos**
Johns Hopkins University

**Paul Smolensky**
Johns Hopkins University
Microsoft Research

**Roland Fernandez**
Microsoft Research

**Jianfeng Gao**
Microsoft Research

## Abstract

Compositionality is a core property of natural language, and compositional behavior (CB) is a crucial goal for modern NLP systems. The research literature, however, includes conflicting perspectives on how CB should be defined, evaluated, and achieved. We propose a conceptual framework to address these questions and survey researchers active in this area.

We find consensus on several key points. Researchers broadly accept our proposed definition of CB, agree that it is not solved by current models, and doubt that scale alone will achieve the target behavior. In other areas, we find the field is split on how to move forward, identifying diverse opportunities for future research.

## 1 Introduction

Compositionality — the ability to correctly process wholes given the ability to correctly process their parts — is a core property of language (Montague, 1973; Fodor and Pylyshyn, 1988), enabling unbounded expressivity through the "infinite use of finite means" (von Humboldt 1836, as quoted by Chomsky 1965). In the past decade, artificial neural network models of natural language have made impressive progress toward human-like language use; however, it is not clear whether their language use consistently demonstrates human-like compositional behavior, especially during generalization (Lake et al., 2019; Hupkes et al., 2020, 2022). This question has been the subject of considerable debate in the field of natural language processing (NLP), as researchers have proposed diverse methods to model and assess compositionality (Pavlick, 2022; Donatelli and Koller, 2023).

We contribute a conceptual organization of current issues surrounding compositionality in artificial neural network models, and use this framework to survey researchers active in this area. We find consensus (roughly 75%+ concordance) on several crucial points. Researchers broadly agree with our proposed definition of compositional behavior (CB, §2.1). They also agree that CB is not a solved problem: current models do not achieve compositional behavior, and scale alone is unlikely to get us there — a perspective consistent with findings from the recent NLP Metasurvey (Michael et al., 2023).

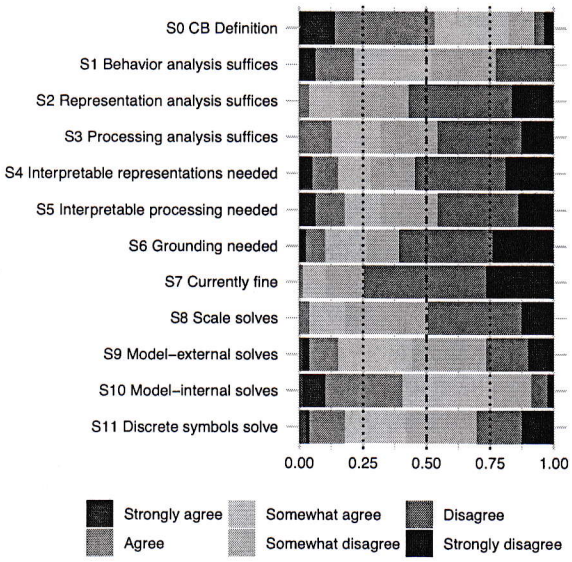


Figure 1: Overview of survey responses. We find consensus (i.e. ~75%+ concordance on "agree" or "disagree") for 7 of the 12 surveyed claims.

In other areas, we find the field is split on how to move forward. In terms of evaluation, researchers disagree on whether current behavioral methods can assess a model's capability for compositional behavior, and remain divided on how best to pursue implementation. We believe there is value for the research community in identifying points of shared understanding and dispute, particularly on a topic foundational to the study of language.

## 2 Framing Compositional Behavior

We conceptually frame our compositionality survey around three key themes, expressed in a series of



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339
November 12-16, 2024 

statements, S0-S11. Respondents provide a graded level of dis/agreement, from Strongly Agree to Strongly Disagree. We first **define** compositional behavior (CB; S0) and ask participants whether they agree with our definition. Given this definition, we then ask which methods are necessary and sufficient to **evaluate** models' capacity for CB (S1-S6). Finally, we ask whether current neural models **achieve** CB (S7), and if not, which interventions are needed (S8-S11).

Here, we briefly review the relevant literature informing each of these sections, and present the corresponding statements in the form that they appear on the survey. Further methodological details of the survey are presented in §3.

### 2.1 Defining Compositional Behavior

Compositionality (Szabó, 2022) has been a topic of extensive debate in the literature on linguistics and philosophy of language. Gottlob Frege is widely recognized as the first philosopher to articulate the concept (Frege, 1914), although his views have been subject to conflicting interpretations (Pelletier, 2001; Herbelot, 2020; Russin et al., 2024). Our goal in this paper is to review the empirical expectations of researchers in computational linguistics, and NLP more broadly; for this reason, our framework focuses on the target *behavior* we would expect a compositional system to exhibit. In so doing, we deliberately sidestep various theoretical and formal distinctions. Here we briefly review our framing of the problem, our proposed definition of compositional behavior, and how it relates to key concepts in the research literature. Many survey participants gave thoughtful and detailed feedback on this definition, which we consider in our discussion (§5).

**Framing the survey** To reduce ambiguity, we asked participants to focus their answers on one particular combination of model and domain. The "current" neural model under consideration is the Transformer and related variants, not including significant changes to the original architecture proposed by Vaswani et al. (2017). The domain under consideration comprises all tasks using natural language (e.g., language modeling, natural language understanding, machine translation, paraphrasing, etc.), formal language (e.g., arithmetic, programming languages, domain-specific languages for specialized tasks such as SCAN and COGS, etc.), or both (e.g., semantic parsing); we exclude other domains such as vision.

**Definition: Compositional Behavior (CB)**

> (**CB**) When a model receives an input $I$ that humans conceive as composed of component parts, if the model produces correct outputs for those parts (in isolation or in other combinations), then it will also produce a correct output for $I$.

Our intended interpretation of (CB) has several key properties. In the following section, sentences in italics were presented to survey participants along with the proposed CB definition.

**Behavior** *CB concerns only behavior, and states nothing about the internal structure or processes of a system or model.* We may consider it situated at Marr's top 'computational' level of analysis (Marr, 1982): CB identifies inputs, outputs, and overall goals, but no particular algorithmic or implementational realization.

**Parts** *CB refers informally to the human conception of inputs and outputs as composed from component parts (conceptual parts, not low-level neural subvector parts), but it does not demand scientific determination of exactly what those parts are.* It does, however, require those parts to be identifiable in more than one context: not only in the input $I$ under consideration, but also in isolation or within another complex expression. The Meaningful Parts Principle (Nefdt, 2020) stipulates that the existence of "meaningful," i.e. composition-relevant, component parts is necessary for any understanding of compositionality. We concur (though see following discussion to clarify "meaning" as distinct from "semantics"), and therefore require identifiable parts to enable CB evaluation. Furthermore, in our stated problem domain of natural and formal language, human-identifiable parts necessarily comprise symbolic sequences and subsequences rather than vector representations.[1]

The broad appeal to human judgment means that CB is not committed to any particular process of linguistic composition. For instance, CB is equally compatible with a bottom-up process which strictly determines a complex expression from its parts (what Pelletier, 2012, calls "building block" compositionality), as with a top-down contextual process which may yield a whole "greater than the sum

[1]Neural network processing is always compositional in the trivial sense that the activation directly resulting from an activation vector is the sum of the activations directly resulting from the subvectors comprising the vector's left and right halves. A useful definition must exclude this trivial sense.

of its parts" (Pelletier's "functional compositionality"). CB also does not require Nefdt's Knowable Parts Principle: the component parts we identify as meaningful for CB evaluation are not required to be similarly meaningful or homomorphic with respect to a model's internal computation. From a practical standpoint, CB is satisfied so long as a human observer deems a model output for input $I$ to be consistent with that same model's outputs for parts of $I$.[2]

**Independence from semantic meaning** *CB does not focus narrowly on the computation of the meaning of expressions; that is merely one case of the highly general phenomenon being targeted.* Compositionality was first developed as a research topic within semantics (Katz and Fodor, 1963), and much current literature reflects this historical focus. For instance, Hupkes et al. (2022) define compositional generalization as a mapping from linguistic input forms to some meaning in a distinct output space, such as in the NLP tasks of semantic parsing or machine translation. They distinguish this from structural generalization occurring entirely within the space of linguistic forms, such as the production of syntactically or morphologically correct sequences. In our proposed definition, however, both of these concepts instantiate compositional behavior. To take a famous example, although the sentence *Colorless green ideas sleep furiously* (Chomsky, 1957) resists truth-conditional semantic interpretation, it recognizably follows the composition structure of English syntax. Another non-linguistic example would be route planning: if a route is known from $X$ to $Y$ and $Y$ to $Z$, CB entails a known route from $X$ to $Z$.

**Independence from learning** *CB does not focus on learning — it states nothing about whether the model has previously encountered input $I$, and only characterizes the target behavior of the model. In a learning context, the type of compositional **generalization** in which the model has not previously seen $I$ is a special case of compositional **behavior*** [bolding added here]. This aspect of CB contrasts with most current literature, which investigates how models might learn to generalize novel input combinations (e.g., Hupkes et al., 2020; Kim and Linzen, 2020). We agree that the generalization scenario presents the key research question; however, defining "generalization" is sufficiently challenging in its own right (e.g., Hupkes et al., 2022). We avoid this challenge by focusing our definition on behavior which covers both known and novel inputs.

[2]Our intended sense of "correct" in the proposed CB definition relies upon human judgment to determine not only the correctness of the input decomposition, but also the correctness of the corresponding outputs; however, only the former is explicitly stated in the definition as written. We discuss this further in §5.

| |
|---|
| **S0**. (CB) is a satisfactory working definition of compositional behavior, an important aspect of compositional generalization. |

Table 1: Survey statement on defining CB (§2.1).

After reading the proposed CB definition and the clarifications above, survey respondents evaluate statement **S0** (Table 1).

## 2.2 Evaluating Compositional Behavior

If we accept the above definition of compositional behavior, which evaluation methods can confirm that a given model is capable of CB? Broadly speaking, there are two main approaches: behavioral and representational. Behavioral evaluation takes a *model-external* view of a system as a black box, relying on carefully designed challenge data and often tightly controlled training data to test performance. Representational evaluation instead focuses on *model-internal* structures and processes. Although researchers often combine behavior and representation analysis in practice, we treat them as distinct here for conceptual clarity.

**Evaluating behavior** In recent years, behavioral evaluation has been used to demonstrate both successes and critical limits in neural models' capacity for compositional generalization. The SCAN dataset (Lake and Baroni, 2018) has been a particularly influential system benchmark (e.g., Dessì and Baroni, 2019; Akyürek et al., 2020; Tan et al., 2020; Newman et al., 2020; Soulos et al., 2020; Kim, 2021; Patel et al., 2022). Like most behavioral challenge sets, SCAN is procedurally generated by a formal language specification. Other notable evaluation datasets generated in this manner include PCFG (Hupkes et al., 2020) to distinguish aspects of combinatory generalization; Colors (Lake et al., 2019; Lake and Baroni, 2023) to compare machine and human few-shot learning; and HANS (McCoy et al., 2019) to address confounds in natural language inference.

While evaluation on formal language data permits fine-grained researcher control, its research implications for natural language performance can

be less clear (cf. Chaabouni et al., 2021). This has motivated the creation of more naturalistic benchmarks to evaluate compositional generalization, such as CFQ (Keysers et al., 2020; Shaw et al., 2021). Though also procedurally generated, COGS (Kim and Linzen, 2020) and recent extensions (Li et al., 2023; Wu et al., 2023) stand out as the most cognitively-motivated benchmarks of this type, with a range of compositional generalization tasks informed by the literature on child language acquisition. Language modeling arguably provides a more cognitively valid objective, but pre-trained language models present further evaluation challenges, as it is difficult to control their exposure (Kim et al., 2022). Survey statement **S1** (Table 2) asks respondents whether the sort of current behavioral evaluation methods reviewed here are *sufficient* to assess a model's capacity for CB.

**Evaluating representations and processing** The external behavior of a model causally depends upon the representations and processes it implements internally. This basic fact has led many researchers to complement behavioral evaluation with model-internal analysis. Pavlick (2023) invokes the classic Chomskyan distinction between competence and performance (Chomsky, 1965) to motivate such approaches, arguing that representation analysis can reveal underlying model capacities (competence) when behavioral evaluation (performance) fails.

There are many techniques to analyze model-internal *representations* (e.g., Belinkov and Glass, 2019; Sajjad et al., 2022; Madsen et al., 2023). One prevalent approach is diagnostic probing (reviewed by Belinkov, 2022), in which an auxiliary model ("probe") is trained to predict certain properties from the internal representations of a main model of interest, thereby indicating how the main model encodes that property. Any representational encoding, however, must be used by model-internal *processes* in order to causally affect the model's behavior. Researchers have explored these causal relations in various ways, such as ablating the representational encodings found by diagnostic probes (e.g., Tucker et al., 2022; Lovering and Pavlick, 2022; Lepori et al., 2023), substituting model components with corresponding interpretable representations (e.g., Soulos et al., 2020; Geiger et al., 2021), and identifying processing circuits associated with particular behaviors (e.g., Olah et al., 2020; Wang et al., 2022; Olsson et al., 2022).

| |
|---|
| **S1**. Current methods for analyzing the behavior of neural models are sufficient to assess whether a model is capable of compositional behavior (CB). For example, consider methods used to assess performance on datasets designed to probe specific aspects of compositional generalization, such as SCAN, COGs, CFQ, PCFG, Colors, etc. |
| **S2**. Current methods for analyzing the representations within neural models are sufficient: if a model is capable of compositional behavior (CB), these analysis methods can identify the model-internal mechanisms supporting this behavior. For example, consider diagnostic probing, visualization, learning interpretable approximations of the representation space, etc. |
| **S3**. Current methods for analyzing the processing within neural models are sufficient: if a model is capable of compositional behavior (CB), these analysis methods can identify the model-internal mechanisms supporting this behavior. For example, consider analysis of circuits / induction heads, causal interventions such as ablation, etc. |
| **S4**. Interpretable representations are necessary: we cannot evaluate whether a model is capable of compositional behavior (CB) unless we can identify human-interpretable parts within its representational structure. |
| **S5**. Interpretable processing is necessary: we cannot evaluate whether a model is capable of compositional behavior (CB) unless we can identify human-interpretable parts within its representational structure, and establish that the model uses these parts as expected during processing. That is to say, if we observe in compositional behavior that certain parts stand in particular relations to one another, we can confirm that those parts interact in similar — ideally isomorphic — ways during the procedure carried out by the model, at some level of description. For example, consider the conceptual roles discussed by Piantadosi and Hill (2022). |
| **S6**. External grounding is necessary: we cannot evaluate whether a model is capable of compositional behavior (CB) unless we can identify human-interpretable parts within its representational structure, and establish that these parts are grounded with respect to some model-external structure in the world. |

Table 2: Survey statements on evaluating CB (§2.2).

While our proposed definition focuses explicitly on compositional *behavior*, one goal of our survey is to assess how researchers in the field view the relationship between internal mechanisms and model performance. Statements **S2–S5** (Table 2) ask whether interpretability in model representations or processing is *necessary* to assess a system's capacity for CB, and whether current methods for evaluating representations or processing are *sufficient* for the same task.

One axis of recent debate has focused on *grounding*: while human language exchanges are grounded (i.e. situated or embedded) in particular social and physical contexts, models of natural

language are exposed only to language. Some researchers (e.g., Bender and Koller, 2020; Bisk et al., 2020) have argued that this lack of grounding seriously impedes language understanding, and Marcus and Murphy (2022) identify this as a key obstacle to compositional generalization. Others (e.g., Piantadosi and Hill, 2022; Santoro et al., 2022; Pavlick, 2023) have argued that, in principle, richly semantically-structured representations can arise through linguistic exposure alone. Statement **S6** (Table 2) asks whether grounded representations are *necessary* to evaluate model capacity for CB.

### 2.3 Achieving Compositional Behavior

Our third set of questions (Table 3) asks whether current models already achieve CB, and if not, how to move forward.[3]

**Non-intervention** The first two statements in this section consider the possibility that we shouldn't worry too much. Perhaps standard architecture modifications and/or pre-training let current models already achieve CB (e.g., Csordás et al., 2021; Ontañón et al., 2022; Lepori et al., 2023; Mueller et al., 2022; Murty et al., 2023; Petty et al., 2024), or perhaps CB will be achieved simply as a byproduct of scale — i.e. given the trajectory of current research. Scale facilitates a wide range of model capabilities (Kaplan et al., 2020; Brown et al., 2020; BIG{-}bench{ }authors, 2023), including compositional generalization (Qiu et al., 2022b); however, the scale paradigm has been criticized (e.g., Linzen, 2020), and the NLP Metasurvey (Michael et al., 2023) reveals widespread skepticism among researchers about scale's potential. Statement **S7** (Table 3) asks respondents whether current models already show sufficient compositional behavior, while **S8** asks whether scale will suffice to attain CB.

**Model-external intervention** The next statement posits that targeted intervention is required, but model-external intervention — i.e. modifications to data and tasks rather than model architecture — will achieve CB. Compositional generalization has been successfully facilitated by approaches such as targeted data augmentation (Andreas, 2019; Akyürek et al., 2020; Qiu et al., 2022a; Patel et al., 2022; Akyurek and Andreas, 2023), auxiliary task supervision (Jiang and Bansal, 2021; Dan et al., 2022), and prompt-tuning (Qiu et al., 2022b; Hahn and Goyal, 2023; An et al., 2023). Statement **S9** (Table 3) asks whether such model-external interventions will suffice.[4]

| |
|---|
| **S7**. Current neural models show a sufficient degree of compositional behavior (CB); we don't need to assign high priority to further research on this topic. |
| **S8**. Current neural models do not show a sufficient degree of compositional behavior (CB), but this issue will likely be resolved as a byproduct of increasing model capacity (i.e. larger models and/or larger datasets). In other words, scale will solve this problem, and we don't need additional interventions to improve compositional behavior. |
| **S9**. Current neural models do not show a sufficient degree of compositional behavior (CB), and some intervention is required, but model-external interventions — as opposed to the model-internal interventions considered in the next claim — are likely to satisfactorily resolve this problem. Examples of model-external interventions include prompt engineering; strategic manipulation or augmentation of training data; and auxiliary tasks during training, pre-training, or fine-tuning. |
| **S10**. Current neural models do not show a sufficient degree of compositional behavior (CB), and model-external interventions are unlikely to resolve this issue. Model-internal interventions or novel architectures, focused on model representations/processing/learning, will be necessary to solve the problem. |
| **S11**. Current neural models do not show a sufficient degree of compositional behavior (CB), and model-internal interventions or novel architectures that incorporate explicit discrete symbolic computation (e.g., program synthesis) will be necessary to solve the problem. |

Table 3: Survey statements on achieving CB (§2.3).

**Model-internal intervention** Statement **S10** (Table 3) posits that novel architectures or other model-internal mechanisms are necessary for CB. Many modeling innovations facilitate compositional generalization, including specialized attention mechanisms (Russin et al., 2019; Li et al., 2019; Korrel et al., 2019; Oren et al., 2020; Bergen et al., 2021), intermediate steps in decoding (Zheng and Lapata, 2021; Ruiz et al., 2021), structured latent variables (Tan et al., 2020; Wang et al., 2021; Herzig and Berant, 2021; Lindemann et al., 2023), and structured distributed representations (Gordon et al., 2020; Smolensky et al., 2022; Soulos et al., 2023). Some

[3] In this section, once respondents answered in the affirmative (i.e. agreed that some approach would solve CB), they could skip later statements.

[4] We note that several effective approaches have paired data-focused interventions with augmented model architectures, such as an auxiliary structure-aware loss function (Yin et al., 2021), memory bank (Lake, 2019), and/or meta-learning objective (Conklin et al., 2021; Lake and Baroni, 2023). We consider such approaches primarily dependent upon the model-external component (e.g., task sampling in the case of meta-learning), and therefore part of this category; however, we note that survey respondents may disagree.

interventions promote compositionality by incorporating discrete symbolic structure, for instance through program synthesis (Nye et al., 2020) or other neuro-symbolic methods (e.g., Chen et al., 2020; Yao and Koller, 2022). Statement **S11** (Table 3) posits the necessity of symbolic computation.

## 3 Survey Methodology

This framework (§2) structures the survey which we circulated to active researchers working in the general area of compositionality, with IRB approval from the University of Edinburgh (RT 541309). The anonymized dataset is available by request for research purposes.

**Distribution** Our target respondent pool for the survey comprised all researchers currently publishing on the topic of compositionality in machine learning. We compiled a list of relevant research publications from three sources:

1. Publications in the ACL anthology[5] since 2019 with "composition", "compositional" or "compositionality" in the title.

2. Publications identified by Hupkes et al. (2023) on the topic of compositional and structural generalization.[6]

3. Publications in prominent machine learning and natural language processing venues (e.g.,, NeurIPS, ICML, ICLR, AAAI, *CL, etc.) which cite Lake and Baroni (2018).[7]

We combined and filtered these three lists, resulting in 246 publications in total.[8] We then extracted all author names with listed contact emails, yielding a contact list of 574 individual researchers.

All of the listed researchers were contacted and invited to participate in the survey, which was open from November 15 to December 15, 2022.[9] We extended further invitations based on personal contacts and the recommendation of other survey respondents, inviting 603 researchers in total. Of these, 57 email addresses were no longer valid, so we assume the invitation reached 546 researchers. 136 (25%) of those researchers opened the link to the survey, and of those, 79 completed the survey. This gives us an overall completion rate of 57% of those who started the survey, representing 13% of the original invitees. While this subsample cannot fully represent the range of views in our target population, we note that these response rates are relatively high with respect to other surveys of educated professionals (Sudman, 1985; Barnhart et al., 2021); for instance, responses to the NLP Metasurvey (Michael et al., 2023) are estimated to cover about 5% of the target demographic.

**Incentives** We invited researchers to contribute their expertise to our survey in a professional capacity; as such, we did not offer any incentives directly to individual respondents. Instead, we committed to donate $10 USD to a charitable organization[10] for each survey completion.[11]

**Survey presentation** Each statement (cf. Tables 1, 2, 3) was presented with the following possible responses: Strongly disagree, Disagree, Somewhat disagree, Somewhat agree, Agree, Strongly agree, with the option to write additional free-form text commentary for each response. Participants gave consent both at beginning of the survey, and at the end, when they were additionally asked to approve use of their name; see Appendix A for details.

**Update period** Our initial data collection period in November 2022 coincided with the release of ChatGPT,[12] followed a few months later by the release of GPT-4 (OpenAI, 2023). These releases received extensive media and public attention, and prompted some research publications reassessing neural models' capacities (e.g., Bubeck et al., 2023). In light of these developments, we offered survey takers the chance to update their responses.

Of the 79 respondents completing the survey, 65 left their email address for follow-up contact. We reached out to these respondents, allowing them to update their original survey responses between July 15 and August 15, 2023. Of the 15 respondents who replied, 10 reported that their views had not changed. Five participants gave updated responses to specific questions, and of those, only two changed their views enough to switch to a dif-

[5] https://aclanthology.org/

[6] https://genbench.org/references

[7] Collected via Semantic Scholar: https://www.semanticscholar.org/

[8] For transparency, we release the list of papers along with further supplementary material at https://github.com/kmccurdy/CBsurvey.

[9] Note that participants could start the survey during this window and finish it at a later time.

[10] Helen Keller International: https://helenkellerintl.org/

[11] Thanks to generous funding from Microsoft, $1000 was ultimately donated.

[12] https://chat.openai.com

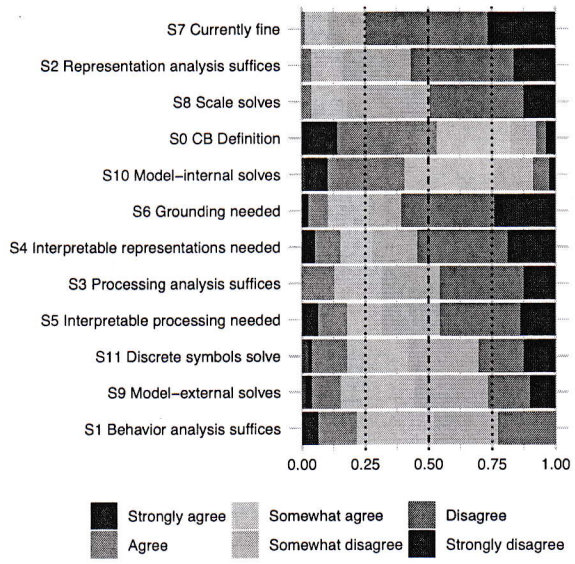


Figure 2: Aggregate survey results ordered from most consensus (top) to most division of opinion (bottom); cf. original presentation order (Figure 1).

ferent cluster (cf. §B).[13] In sum, the respondents to our update message — roughly 20% of survey participants — largely retain their original views. We take this as evidence that the opinions gathered in the survey remain representative, recent technological developments notwithstanding.

## 4 Survey Results

Aggregate responses are shown in Figure 2 (see also Figure 1 for presentation in survey order). To our surprise, we found much more agreement across the community than expected, with researchers expressing a consensus opinion for 7 of the 12 statements listed on the survey.

We define "consensus" as survey statements for which roughly 75% or more of respondents converge on agreement (i.e. Strongly agree, Agree, or Somewhat agree) or disagreement. 81% of respondents **agree** with the statement S0, our proposed definition of Compositional Behavior (CB). We also find near-consensus agreement for statement S10: 73% of respondents agree that model-internal interventions are likely necessary to achieve CB.

Otherwise, we found consensus on **disagreement**. On the topic of interpretable representations, 82% of respondents judge that current methods are not sufficient to evaluate CB (S2), but 74% also judge that interpretable representations are not necessary for this evaluation (S4), and 75% do not find grounded representations necessary (S6). On the topic of achieving CB, 88% of respondents agree that current models do not achieve CB (S7), and 81% do not think it will be achieved by scale (S8).

The scale result mirrors findings from the larger NLP Metasurvey (Michael et al., 2023, 16336): 83% of their 327 respondents disagree with the view that scaling up would solve "practically any important problem in NLP," and 71% believe that NLP research is excessively focused on scale. We interpret these convergent findings as evidence that skepticism about scale is not restricted to researchers focused on compositionality, but characteristic of the broader NLP community.

**Points of division** Some statements show a near even split of opinion. Researchers are divided on the adequacy of current behavioral methods to evaluate CB (S1), with 53% finding them acceptable. Opinions also differ on how to achieve CB; 43% think that model-external interventions will be sufficient (S9), but 40% consider discrete symbolic structure necessary (S11).

To better represent fine-grained differences in opinion, we performed principal component analysis. Figure 3 visualizes the two main axes of variation in responses: on the necessity of interpretable processes and representations, and on the adequacy of current methods — especially behavioral methods — for evaluating CB. We additionally identified respondents with one of six clusters, ordered from largest to smallest: Default View, Minimal Interventionist, Current Analysis Suffices, Grounded Symbolic Interpretability, Minimal Interpretability, and Non-interventionist. For details on the cluster analysis, see Appendix B.

## 5 Discussion

Beyond the quantitative overview in §4, many survey respondents provided highly thoughtful written comments. We regret our inability to thoroughly engage all of the excellent points raised. Here, we discuss three key statements — our proposed definition of CB (S0), the adequacy of behavioral evaluation (S1), and the need for interpretable representations (S4) — in light of the nuanced perspectives found in the comments. We focus on the role of model interpretability and the adequacy of current evaluation because these concepts roughly correspond to the main axes of variation identified in our principal components analysis (Figure 3).

[13] For the five participants who updated their responses, Figure 3 plots their new position and cluster assignment.

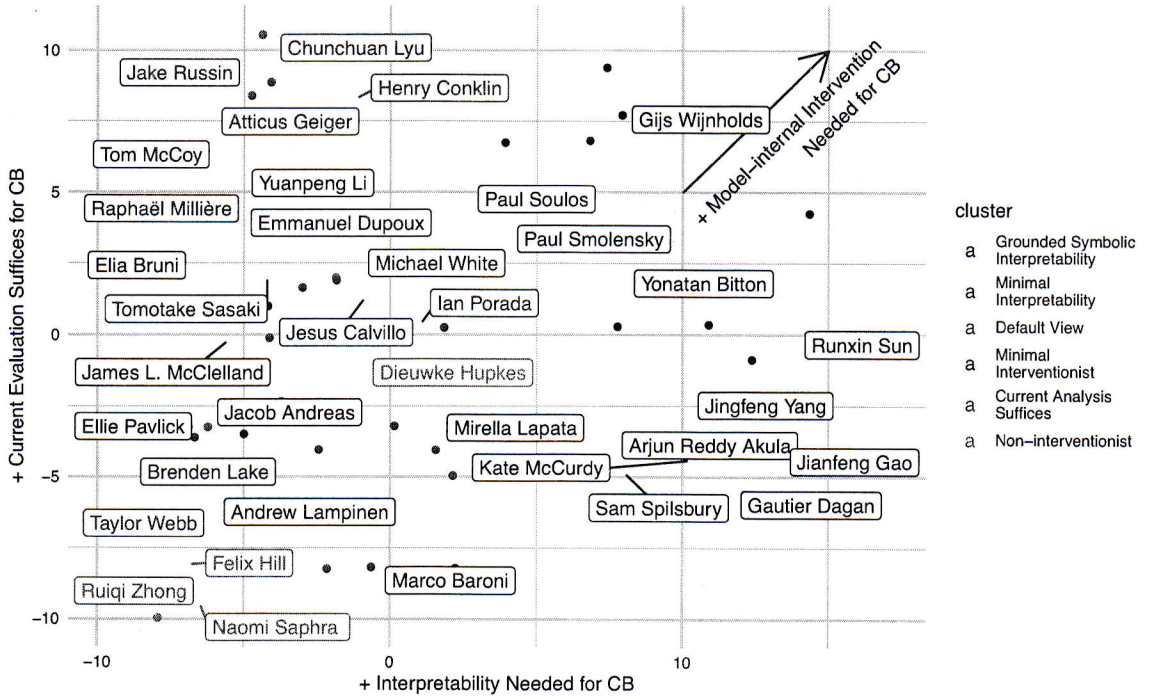


Figure 3: Logical geography of CB survey responses, inspired by Dennett (1986). Individual respondents are projected to a two-dimensional location using principal component analysis and colored based on cluster (cf. Appendix B). Points represent participants who did not give permission to use their name. Axis labels reflect our loose interpretation of the principal components.

**Defining CB** As discussed in §2.1, we propose defining compositional behavior (CB) with respect to an informal human-like conception of wholes and parts. Though a few commenters consider such human-level perceptibility an irrelevant constraint, most agree with the criterion; many of those who agree, however, nonetheless find CB too vague, or insufficiently formal for useful research. Several highlight the difficulty of finding consensus in human judgments. For instance, Andrew Lampinen cites Gleitman and Gleitman (1971)'s finding that educational level affects semantic composition in compound words, and Lake and Baroni (2023) observe considerable variability in the composition rules used by human participants in a highly constrained experimental setting. We recognize the diverse nature of human judgment, and the challenge for scientific evaluation.

We also thank respondents for highlighting an overlooked ambiguity in our proposed CB definition: while we intend our appeal to human judgment to apply to both a) the decomposition of an input $I$ into parts and b) the correctness of the respective model output, the definition as written only states (a) explicitly. Dieuwke Hupkes, James L. McClelland, and Andrew Lampinen each propose amended CB definitions which directly incorporate (b). Many other comments raise related points: that correct decomposition of the input does not entail correct composition of the output, that decomposition and composition are contextually variable in natural language, that partial composition is possible, and that the meaning of composed expressions in natural language often rely upon factors beyond the contents of input parts. We find these observations insightful, and consider them at least partly addressed by deferring to human judgment of compositional outcomes, despite the challenges outlined above.

A final key point raised by several commenters is the central importance of generalization. Ellie Pavlick, Jake Russin, Dieuwke Hupkes, and Emmanuel Dupoux, inter alia, note that a model which achieves compositional behavior on a given dataset through memorization would not be interesting from a research perspective, as we would not expect it to extend CB to other datasets and domains. This contention highlights the central challenge of CB evaluation for machine learning: how can we be sure that compositional behavior arises for the right reasons?

**Behavioral evaluation** Survey respondents are almost perfectly divided on the adequacy of current methods for behavioral evaluation. 53% agree that current behavioral methods are sufficient to establish CB — though as noted by Raphaël Millière and others, this requires proper experimental design: careful control of training data, such that the model is not exposed to the generalizations necessary to succeed on the test set. Behavioral evaluation also permits greater ecological validity, as we can often directly compare human performance on the same behavioral task.

The other 47% of respondents are more skeptical. Marco Baroni and Andrew Lampinen characterize current behavioral methods as "necessary, but not sufficient;" many other commenters note that behavioral evaluation on a limited phenomenon or domain cannot establish fully general CB, and raise concerns about synthetic tasks which may not reflect performance in more realistic settings. We note that many respondents who agree with S1 nonetheless raise similar concerns in their comments. Researchers have a shared view of the limitations of current behavioral evaluation, but differ on whether these limitations prevent these methods from sufficiently demonstrating CB.

**Interpretable representations** We were particularly interested in how researchers view the connection between interpretable representations and evaluating compositional behavior (CB). The results reveal a strong consensus that no such connection is necessary. Of those disagreeing with statement S4 (Table 2), many commenters note that CB is behavioral by definition, hence model-agnostic behavioral evaluation must suffice in principle, and many additionally observe that we rely on behavioral rather than representational evidence of compositionality in humans. Generalization is important here as well: several commenters note that full data coverage of the relevant domain is required for behavioral evaluation to adequately demonstrate CB. Many of those who disagree with S4 nonetheless affirm scientific interest in representational structure, and consider interpretable representations informative and helpful, if not required, for CB evaluation. Among the minority who find interpretable representations necessary, comments emphasize the need for formal verification of the mechanisms supporting CB, and the inadequacy of behavioral evaluation in this regard.

**Toward compositional behavior** Based on the perspectives reviewed here, we see several practical implications for future research. First, there is substantial room for progress in the domain of **interpretability**, as a majority of respondents find current approaches inadequate (S2). Even though most also reject the idea that interpretability is *necessary* to establish CB (S4, S5), many comments clarify that interpretability is still desirable for scientific purposes (cf. Mosbach et al., 2024), and can help us distinguish fundamental limitations in model capabilities from performance failures driven by other issues (Pavlick, 2023). Second, a key finding of our survey is that most researchers consider human behavior an acceptable reference for defining correct compositional behavior (S0), but differ on whether current behavioral evaluation methods are satisfactory (S1). This suggests that **behavioral evaluation** could be improved. Respondents identify diverse approaches such as directly comparing human and model performance (e.g., Lake and Baroni, 2023; Lampinen, 2022), developing more naturalistic tasks, and evaluating on a broader range of domains. We note a certain duality in evaluation: establishing CB requires detailed knowledge of *either* model-internal workings (to verify compositional capacities; e.g., Lepori et al., 2023) *or* the full set of training data (to rule out learning non-compositional shortcuts; e.g., Hupkes et al., 2022). Third, we see considerable diversity of opinion in terms of **modeling interventions** to achieve CB. While most researchers are skeptical of scale (S8) and expect internal changes to model architectures (S10), half of respondents think CB can be achieved through model-external approaches (S9), but the other half think that model-internal symbolic processing is likely required (S11). Many avenues for exploration remain open; above all, respondents strongly agree that the problem of CB is not yet solved (S7).

## 6 Conclusion

Compositionality, a foundational aspect of natural language, has taken on new significance in light of modern neural models and uncertainty about their capacities. This paper offers a framework for defining, evaluating, and achieving compositional behavior in neural models, and surveys the views of researchers active in this area. We identify key points of consensus and division, providing a snapshot of the field to inform future research.

## Acknowledgments

We offer our heartfelt thanks to all participants in our survey for their consideration and expertise, to the University of Edinburgh School of Informatics for hosting the survey and providing institutional ethics review, and to Microsoft Research for generously funding the survey incentive donations. The first author is funded by the Deutsche Forschungsgemeinschaft (DFG Project-ID 232722074, SFB 1102), and worked on this project as an intern at Microsoft Research and doctoral student at the University of Edinburgh.

## Limitations

There some potentially critical conceptual limitations to our approach. One limitation of our survey is the fact that all later statements rely upon acceptance of the first statement, namely our proposed definition of CB; therefore, conceptual issues in this definition may affect the validity of the entire survey. In the discussion section (§5), we consider some issues with our wording of the CB definition, along with proposed amendments raised by survey respondents. Another possible objection is that our proposed CB definition is too broad, and insufficiently specified to elicit meaningful disagreement within the research community. We do not entirely agree with this objection, as we consider having a shared if underspecified working definition to be valuable in its own right; however, we acknowledge that this breadth may limit the scientific contribution of this work. Finally, we deliberately limited the architecture under consideration to the Transformer, and the domain under consideration to natural and formal languages, even though compositional behavior is also important in other areas of NLP and AI.

A second set of limitations is methodological. While we attempted to include a diverse range of perspectives from the field, including senior and junior researchers, our survey sample cannot be perfectly representative and a different recruitment method may have yielded different results. Another consideration is in the use of respondents' names: while we strove to follow best ethical practices in this regard (see Appendix A), some may still raise objections to our use of respondents' names in this paper. Finally, a substantial limitation of this paper submission format is that we have not had the space to fully engage with the many, many thoughtful and detailed responses shared by survey participants. We deeply appreciate the time and energy that respondents spent on this survey, and regret our inability to give all the responses the attention they merit.

## References

Ekin Akyurek and Jacob Andreas. 2023. LexSym: Compositionality as Lexical Symmetry. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 639–657, Toronto, Canada. Association for Computational Linguistics.

Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. 2020. Learning to Recombine and Resample Data for Compositional Generalization. *arXiv:2010.03706 [cs]*.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How Do In-Context Examples Affect Compositional Generalization? ArXiv:2305.04835.

Jacob Andreas. 2019. Good-Enough Compositional Data Augmentation. *arXiv:1904.09545 [cs]*.

Brendan J. Barnhart, Siddharta G. Reddy, and Gerald K. Arnold. 2021. Remind Me Again: Physician Response to Web Surveys: The Effect of Email Reminders Across 11 Opinion Survey Efforts at the American Board of Internal Medicine from 2017 to 2019. *Evaluation & the Health Professions*, 44(3):245–259.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Leon Bergen, Timothy O'Donnell, and Dzmitry Bahdanau. 2021. Systematic generalization with edge transformers. *Advances in Neural Information Processing Systems*, 34.

BIG{-}bench{ }authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr

Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. Arxiv:2004.10151.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs].*

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN. ArXiv:2107.01366.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. 2020. Compositional Generalization via Neural-Symbolic Stack Machines. In *Advances in Neural Information Processing Systems*, volume 33, pages 1690–1701. Curran Associates, Inc.

Noam Chomsky. 1957. *Syntactic structures*. Mouton de Gruyter, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. 2021. Meta-Learning to Compositionally Generalize. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3322–3335, Online. Association for Computational Linguistics.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Soham Dan, Osbert Bastani, and Dan Roth. 2022. Understanding Robust Generalization in Learning Regular Languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4630–4643. PMLR.

Daniel C. Dennett. 1986. The Logical Geography of Computational Approaches: A View From the East Pole. In Myles Brand and Robert M. Harnish, editors, *The Representation of Knowledge and Belief*. University of Arizona Press.

Roberto Dessì and Marco Baroni. 2019. CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. *arXiv:1905.08527 [cs].*

Lucia Donatelli and Alexander Koller. 2023. Compositionality in Computational Linguistics. *Annual Review of Linguistics*, 9(Volume 9, 2023):463–481. Publisher: Annual Reviews.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gottlob Frege. 1914. Letter to Jourdain. *Philosophical and mathematical correspondence*, pages 78–80. Publisher: Chicago University Press.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

L Gleitman and H Gleitman. 1971. *Phrase and paraphrase: Some innovative uses of language*. Norton Press, New York.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. 2020. Permutation Equivariant Models for Compositional Generalization in Language.

Michael Hahn and Navin Goyal. 2023. A Theory of Emergent In-Context Learning as Implicit Structure Induction. ArXiv:2303.07971.

Aurelie Herbelot. 2020. How to Stop Worrying About Compositionality. *The Gradient.*

Jonathan Herzig and Jonathan Berant. 2021. Span-based Semantic Parsing for Compositional Generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921, Online. Association for Computational Linguistics.

Wilhelm von Humboldt. 1836. *Über die Verschiedenheit des Menschlichen Sprachbaues.*

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann,

Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review. ArXiv:2210.03050.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174. Publisher: Nature Publishing Group.

Yichen Jiang and Mohit Bansal. 2021. Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*.

Jerrold J. Katz and Jerry A. Fodor. 1963. The Structure of a Semantic Theory. *Language*, 39(2):170.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. *arXiv:1912.09713 [cs, stat]*.

Najoung Kim and Tal Linzen. 2020. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled Lexical Exposure Leads to Overestimation of Compositional Generalization in Pretrained Models. ArXiv:2212.10769.

Yoon Kim. 2021. Sequence-to-Sequence Learning with Latent Neural Grammars. *arXiv:2109.01135 [cs]*.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. 2019. Transcoding Compositionally: Using Attention to Find More Generalizable Solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR. ISSN: 2640-3498.

Brenden M. Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. *arXiv:1906.05381 [cs]*.

Brenden M. Lake and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121. Number: 7985 Publisher: Nature Publishing Group.

Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. ArXiv: 1901.04587.

Andrew Kyle Lampinen. 2022. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans.

Michael Lepori, Thomas Serre, and Ellie Pavlick. 2023. Break It Down: Evidence for Structural Compositionality in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 42623–42660. Curran Associates, Inc.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A Structural Generalization Benchmark for Semantic Parsing. ArXiv:2310.15040.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019. Compositional Generalization for Primitive Substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4293–4302, Hong Kong, China. Association for Computational Linguistics.

Matthias Lindemann, Alexander Koller, and Ivan Titov. 2023. Compositional Generalisation with Structured Reordering and Fertility Layers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2172–2186, Dubrovnik, Croatia. Association for Computational Linguistics.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics. ArXiv: 2005.00955.

Charles Lovering and Ellie Pavlick. 2022. Unit Testing for Concepts in Neural Networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 55(8):1–42.

Gary Marcus and Elliot Murphy. 2022. Three ideas from linguistics that everyone in AI should know.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press. Google-Books-ID: D8XxCwAAQBAJ.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv:1902.01007 [cs]*.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Richard Montague. 1973. The Proper Treatment of Quantification in Ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, Synthese Library, pages 221–242. Springer Netherlands, Dordrecht.

Marius Mosbach, Vagrant Gautam, Tomás Vergara-Browne, Dietrich Klakow, and Mor Geva. 2024. From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2023. Characterizing Intrinsic Compositionality in Transformers with Tree Projections. arXiv.

Ryan M. Nefdt. 2020. A Puzzle concerning Compositionality in Machines. *Minds and Machines*, 30(1):47–75.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. 2020. The EOS Decision and Length Extrapolation. *arXiv:2010.07174 [cs]*.

Maxwell Nye, Armando Solar-Lezama, Josh Tenenbaum, and Brenden M Lake. 2020. Learning Compositional Rules via Neural Program Synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 10832–10842. Curran Associates, Inc.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. Making Transformers Solve Compositional Tasks. ArXiv:2108.04378.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving Compositional Generalization in Semantic Parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, Phil Blunsom, and Navin Goyal. 2022. Revisiting the Compositional Generalization Abilities of Neural Sequence Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 424–434, Dublin, Ireland. Association for Computational Linguistics.

Ellie Pavlick. 2022. Semantic Structure in Deep Learning. *Annual Review of Linguistics*, 8(Volume 8, 2022):447–471. Publisher: Annual Reviews.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

Francis Jeffry Pelletier. 2001. Did Frege Believe Frege's Principle? *Journal of Logic, Language and Information*, 10(1):87–114.

Francis Jeffry Pelletier. 2012. *Holism And Compositionality*. Oxford University Press.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The Impact of Depth on Compositional Generalization in Transformer Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252, Mexico City, Mexico. Association for Computational Linguistics.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. ArXiv:2208.02957.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Paweł Krzysztof Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. 2022a. Improving Compositional Generalization with Latent Structure and Data Augmentation. ArXiv:2112.07610.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022b. Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9157–9179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Luana Ruiz, Joshua Ainslie, and Santiago Ontañón. 2021. Iterative Decoding for Compositional Generalization in Transformers. ArXiv:2110.04169.

Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. 2024. From Frege to chatGPT: Compositionality in language, cognition, and deep neural networks.

Jake Russin, Jason Jo, Randall C. O'Reilly, and Yoshua Bengio. 2019. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv:1904.09708 [cs, stat]*.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level Interpretation of Deep NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. 2022. Symbolic Behaviour in Artificial Intelligence. ArXiv:2102.03406.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both? *arXiv:2010.12725 [cs]*.

Paul Smolensky, Richard McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. 2022. Neurocompositional Computing: From the Central Paradox of Cognition to a New Generation of AI Systems. *AI Magazine*, 43(3):308–322. Number: 3.

Paul Soulos, Edward Hu, Kate McCurdy, Yunmo Chen, Roland Fernandez, Paul Smolensky, and Jianfeng Gao. 2023. Differentiable Tree Operations Promote Compositional Generalization. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Paul Soulos, Tom McCoy, Tal Linzen, and Paul Smolensky. 2020. Discovering the Compositional Structure of Vector Representations with Role Learning Networks. ArXiv:1910.09113.

Seymour Sudman. 1985. Mail Surveys of Reluctant Professionals. *Evaluation Review*, 9(3):349–360. Publisher: SAGE Publications Inc.

Zoltán Gendler Szabó. 2022. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, fall 2022 edition. Metaphysics Research Lab, Stanford University.

Shawn Tan, Yikang Shen, Timothy J. O'Donnell, Alessandro Sordoni, and Aaron Courville. 2020. Recursive Top-Down Production for Sentence Generation with Latent Trees. ArXiv: 2010.04704.

Mycal Tucker, Tiwalayo Eisape, Peng Qian, Roger Levy, and Julie Shah. 2022. When Does Syntax Mediate Neural Language Model Performance? Evidence from Dropout Probes. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5393–5408, Seattle, United States. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021. Structured Reordering for Modeling Latent Alignments in Sequence Transduction. In *Advances in Neural Information Processing Systems*, volume 34, pages 13378–13391. Curran Associates, Inc.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. ArXiv:2211.00593.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation. ArXiv:2303.13716.

Yuekun Yao and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengcheng Yin, Hao Fang, Graham Neubig, Adam Pauls, Emmanouil Antonios Platanios, Yu Su, Sam Thomson, and Jacob Andreas. 2021. Compositional Generalization for Neural Semantic Parsing via Span-level Supervised Attention. In *Proceedings of the 2021 Conference of the North American Chapter of*

*the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Online. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2021. Compositional Generalization via Semantic Tagging. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1022–1032, Punta Cana, Dominican Republic. Association for Computational Linguistics.

## A Consent and Data Use

Our survey is somewhat unusual in that our target population comprises researchers who have published on a particular topic. Therefore, naming specific individuals and their opinions can be viewed as part of the broader scientific project; nevertheless, personally identifiable data requires sensitive handling even for purposes in the public interest. We address this by requesting consent at three different points in the survey process.

**Initial consent** Before taking the survey, each participant read an IRB-approved (RT 541309, University of Edinburgh School of Informatics) information sheet on the goals and contents of the study, data protection measures, and contact information. In order to proceed to the survey, each participant approved the following statement:

> By proceeding with the study, I agree to all of the following statements:
>
> - I have read and understood the above information.
> - I understand that my participation is voluntary, and I can withdraw at any time.
> - I consent to my anonymised data being used in academic publications and presentations.
> - I allow my data to be used in future ethically approved research.

**Retrospective consent** At the end of survey, we asked participants to provide a more detailed form of consent, including use of their name. We reasoned that, after seeing the contents of the survey, participants would be better able to make an informed decision on choosing whether to be named. Participants were asked to select one of the following options:

> Please indicate which uses of your data you consent to.
>
> - I consent to the analysis and release of my anonymized data, and you can use my name to quote my written answers.
> - I consent to the analysis and release of my anonymized data, and you can anonymously quote my written answers.
> - I consent to the analysis and release of my anonymized data, but please do not quote my written answers.
> - I do not consent to any use, please do not include my data in your analysis.

**Update clarification** Following the initial round of responses, we reached out to survey participants during an update round as described in §3. In this follow-up communication, we included the original survey responses provided by each individual participant, and a brief description of the cluster analysis. We also attached a draft version of Figure 3 with the participant's name included, if they consented to use of their name, or anonymized if they had not. We clarified to participants that they had the option to revoke use of their name if they did not wish to appear on the plot — or, conversely, they could approve use of their name on the plot if they had previously opted for anonymity. At this stage, one participant revoked use of their name, and one participant granted it.

## B Cluster analysis

We performed unsupervised hierarchical clustering on the responses using the `hclust` method in R (R Core Team, 2023). Responses were transformed to a numerical scale and additionally adjusted to strongly differentiate agreement from disagreement, yielding a range from 3.5 to 5.5 on the positive side, and $-5.5$ to $-3.5$ on the negative. We used the "complete linkage" clustering method, which computes proximity across clusters using the most distant instances ("furthest neighbors"), thereby minimizing the upper-bound distance between members of the same cluster. We found that the 6-cluster grouping explained 90% of the variance across responses, and increasing the cluster count did not produce notable improvements. Figure 4 shows the distribution of responses within each cluster and Figure 3 projects each individual participant to a two-dimensional plane using principal component analysis. Here, we describe the

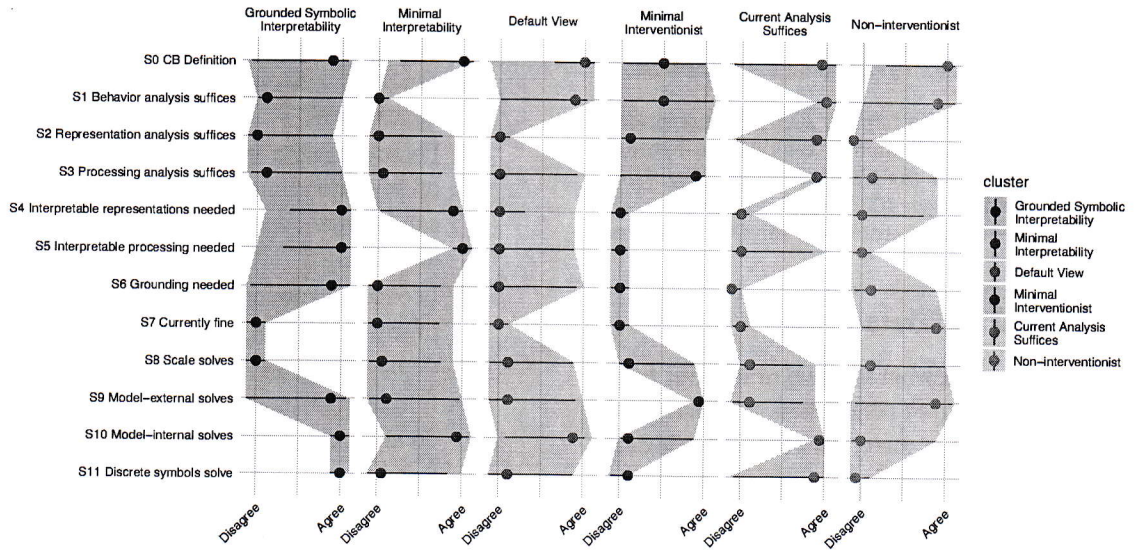


Figure 4: Distribution of responses for each cluster. Point shows the median response on a transformed scale, line shows 95% distribution tail, shaded area shows full range of responses per cluster. Clusters are ordered roughly based on their centroid projection on the first principal component; the reverse order is shown on Figure 3's x-axis.

six clusters ordered from largest to smallest.

**Default View** The largest cluster, comprising 29% of respondents, reflects what we call the "default" position. Like the majority of survey participants, members of the Default View cluster agree with our proposed definition of CB (S0), find CB in current models insufficient (S7), and do not consider the analysis of interpretable representations currently adequate (S2) or necessary (S5) to evaluate CB. While they show a broader range of views on other statements, the central tendency of this cluster typically reflects majority opinion. We describe the following clusters in terms of how they deviate from the Default View.

**Minimal Interventionist** Compared to the Default View, the Minimal Interventionist position (18%) largely doubts that model-internal interventions (S10) are needed to achieve CB, and sees model-external interventions (S9) as sufficient. Unlike Non-interventionists, however, they still see CB as an open problem (S7). This cluster is also strongly committed to the majority stance that interpretable (S4, S5) and grounded (S6) representations are not needed for CB evaluation, and inclined to favor current analysis methods for processing (S3). Finally, compared to other clusters, members of this cluster are most likely to disagree with our proposed definition of CB (S0).

**Current Analysis Suffices** Respondents in this cluster (15%) find that current analysis methods are sufficient across the board: for behavior (S1), representations (S2), and especially processing (S3). They are also united on the need for model-internal interventions to achieve CB (S10), and the lack of necessity for interpretable (S4) or grounded (S6) representations in evaluation.

**Grounded Symbolic Interpretability** These respondents (15%) are committed to the need for symbolic internal modifications of models (S11), and decisively reject scale as a solution (S8). They also find interpretability necessary in both representations (S4) and processing (S5), and are most likely to favor grounded representations (S6).

**Minimal Interpretability** Contrary to the Default View, the Minimal Interpretability position (11% of respondents) identifies interpretable processing (S5) as critical for CB evaluation, and favors interpretable representations (S4). They share this view with the Grounded Symbolic Interpretabilty position, but differ in rejecting the need for grounding (S6) and discrete symbolic structure (S11). This cluster also firmly rejects the adequacy of current behavioral methods to evaluate CB (S1).

**Non-interventionist** Respondents in the smallest cluster (8%) are most likely to view current models as already achieving adequate CB (S7). They consider external interventions (S9) sufficient to

handle any remaining issues, with no likely need for internal modifications (S10), especially symbolic computation (S11).